CHAPTER 11

Word alignment in the Russian-Chinese parallel corpus

Anastasia Politova[1], Olga Bonetskaya[2], Dmitry Dolgov[3],
Maria Frolova[3] & Anna Pyrkova[2]
[1] Soochow University | [2] HSE University | [3] Independent researcher

The Russian-Chinese parallel corpus (RuZhCorp) was created in 2016 by sinologists and computational linguists. So far, it has accumulated 1 074 texts and over 4.6 million words that are aligned on a sentence level. To produce word alignment for the entire corpus, we used deep neural networks trained both on the whole RuZhCorp and on a manually aligned at a word level gold dataset. Using the principles presented in previous publications, we compiled the first word-to-word alignment guideline for the Russian-Chinese language pair, which makes the manual alignment process less ambiguous and more consistent. The joint fine-tuning of the LaBSE deep learning model on RuZhCorp and the gold dataset achieved the best AER of 18.9%.

Keywords: word alignment, gold dataset, linguistic guideline, deep learning, language model

1. Introduction

Word alignment of parallel corpora is defined as finding word-to-word[1] relationships between bitexts already aligned on a sentence level (Brown et al. 1990). A fundamental task in both natural language processing (NLP) and linguistics, word alignment not only serves as a basis for further research or information extraction in multilanguage search systems (Davis & Dunning 1995; Nie et al. 1999; Chen & Nie 2000) but can also be considered a final product for end users: instead of regular dictionaries or machine translation tools, people may use

[1] By "word-to-word" relations, we also mean the relations on the level of lexical units, where lexical units may be represented by single words or groups of words. This approach allows to align verb phrases and idiomatic expressions as single slots with one-to-many or many-to-many relations.

https://doi.org/10.1075/scl.113.11pol

context-aware dictionaries that show how a given word has been translated by real human translators in different contexts.[2] Word-aligned corpora are beneficial for bilingual lexical or grammar usage extraction (Kuhn 2004) and assist linguists and professionals working with language data (translators, foreign language teachers, etc.) in theoretical and language data search (Östling 2016; Wälchli & Cysouw 2012; Mayer & Cysouw 2012; Cysouw & Wälchli 2007).

In NLP, word alignment is intricately related to machine translation (MT). Historically, automatic word alignment per se has mainly been done using statistical methods. The expectation-maximization algorithm was first proposed by Dempster et al. (1977) and implemented for word alignment under the name of IBM models by Brown et al. (1993). Och and Ney (2003) created a tool called GIZA+ + that remains a standard benchmark until now. Later, several MT-related approaches were proposed. Bahdanau et al. (2015) used a DNN (deep neural network) to learn to align and translate jointly. Stengel-Eskin et al. (2019) used supervised learning to extract alignments from the attention module of a Transformer DNN. Several papers using BERT pre-trained models were published in recent years (Dou & Neibig 2021; Nagata, Chousa & Nishino 2020; Li et al. 2019). Some authors conversely used alignment data to either improve or explain machine translation (Chen et al. 2016; Tamer & Ney 2017; Stahlberg, Saunders & Byrne 2018). All the mentioned models are mainly trained on English plus one other language.

However, the results of existing aligners still need to be evaluated and further developed for other language pairs. Previous research reveals that most models show better AER (Alignment Error Rate) when trained on a manually annotated gold dataset. So far, there are several manually annotated gold datasets: Myanmar-English (Han & Thida 2019), Hindi-English (Yadav & Gupta 2010), Dutch-English (Macken 2010), Chinese-Korean (Li, Kim & Lee 2008), six pairs of 4 cognate languages (Graça et al. 2008), Czech-English (Kruijff-Korbayová, Chvátalová & Postolache 2006), English-Spanish (Lambert 2005), and English-French (Och & Ney 2003). However, no gold dataset is available for Russian-Chinese word alignment. Therefore, the current study expands the language pairs list, presents the first word-alignment manual for Russian-Chinese, and provides a deep learning language model trained on the Russian-Chinese parallel corpus (RuZhCorp, < https://linghub.ru/rnc_parallel_chinese/search >).

This collaborative paper written by linguists and data scientists presents our results for RuZhCorp word alignment with a neural network model. Section 2 describes how the gold dataset of Russian-Chinese sentence pairs was created and

[2] Several decades earlier word aligned corpora were used for automatic dictionary and concordance lists compilation (Sahlgren & Karlgren 2005).

elaborates on the developed rules for Russian-Chinese alignment. This should be helpful for further improvements of word alignment models for this pair of languages and any other. Section 3 presents results obtained at different stages of model training and compares them to the results for other comparable pairs of non-similar languages. The final section concludes with our results and discusses the significance and future perspectives of our work.

2. Corpus

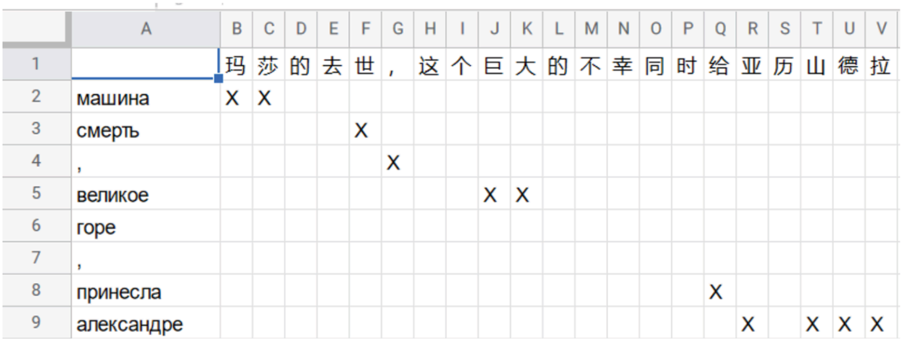
2.1 Building the gold dataset

2.1.1 *Types of alignment*

A gold dataset is a set of sentences manually aligned by linguists according to pre-established guidelines that make the alignment process as unambiguous as possible. Following the experience of Graça et al. (2008) and Och and Ney (2000), who both differentiated between Sure and Possible alignments, we similarly distinguished between S(ure) and P(ossible) alignments. However, in comparison to the previous works, our demarcations of S and P are slightly different. Och and Ney used S-alignment for unambiguous alignments and P(ossible) for those that might or might not exist. Besides, "the P relation is used especially to align words within idiomatic expressions, free translations, and missing function words" (Och & Ney 2000). For Graça et al., S-alignments represent a translation that is possible in every context, and "P-alignments when translation [is] possible in certain contexts or in the presence of functional words might be absent in one of the languages of a language pair" (Graça et al. 2008). We define S-alignments as a sure/direct/unarguable word translation. S-alignments are also used to align the established idiomatic expressions that are marked as many-to-many sure correspondences. P-alignment is for a translation that is possible in a certain context or is a euphemistic translation of a word. In other words, our alignment is paradigmatic and not syntagmatic, that is words/expressions are aligned based on meaning or usage differences and not on the grammatical structure (see Section 2.2 and Figure 3).

2.1.2 *Alignment tool*

Manual word alignment requires a tool where the source and target sentences are presented intuitively and compactly. Graça et al. (2008) used a special software called Alignment Tool. We used Google Sheets to display source and target sentences in the same way as they were shown in the Alignment Tool: source and target sentences arranged in column and row headers of a spreadsheet with each word put in a different cell. One sheet or tab represents one sentence pair only. In order to minimize the manual work and provide annotators with a draft, the body of the table was prepopulated with the results of a word-to-word alignment performed by an unsupervised learning model, where “X” represents the matches made by the system. Figure 1 shows our raw input. Such a simple tool is more accessible and allows the work to be done simultaneously by several linguists working remotely. Additionally, Google Sheets easily shows what modifications have been made recently.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		玛	莎	的	去	世	，	这	个	巨	大	的	不	幸	同	时	给	亚	历	山	德	拉
2	машина	X	X																			
3	смерть					X																
4	,						X															
5	великое									X	X											
6	горе																					
7	,																					
8	принесла																X					
9	александре																	X		X	X	X

Figure 1. Working page from Google Sheets with the Chinese sentence in column headers and the Russian sentence in row headers; the “X” in the matching area is the result of word-to-word alignment produced by an unsupervised learning model

2.1.3 *The alignment process*

At the start of the project, we aligned 125 pairs of Chinese sentences translated into Russian in a full manual mode (without using the aforementioned unsupervised learning model) and then applied the described process to another 327 pairs, bringing the total size of the gold dataset to 452 pairs.

Due to a limited number of annotators and time and resource constraints, we chose the iterative way (that is a first annotator manually aligns tokens by changing the automatic alignment, followed by the second annotator, who verifies the manual alignment and, in case of disagreement or doubts, puts the issue for the discussion) instead of having two annotators working parallelly on the same data. Consequently, based on the existing alignment guidelines, we developed our own alignment flow (see Figure 2); and four annotators got spreadsheets similar to the one shown in Figure 1.

In other words, our alignment process is structured in such a way as to make it transparent and objective, that is based on rules rather than on a personal opinion of a linguist-aligner. Disputed matches were discussed until an agreement on

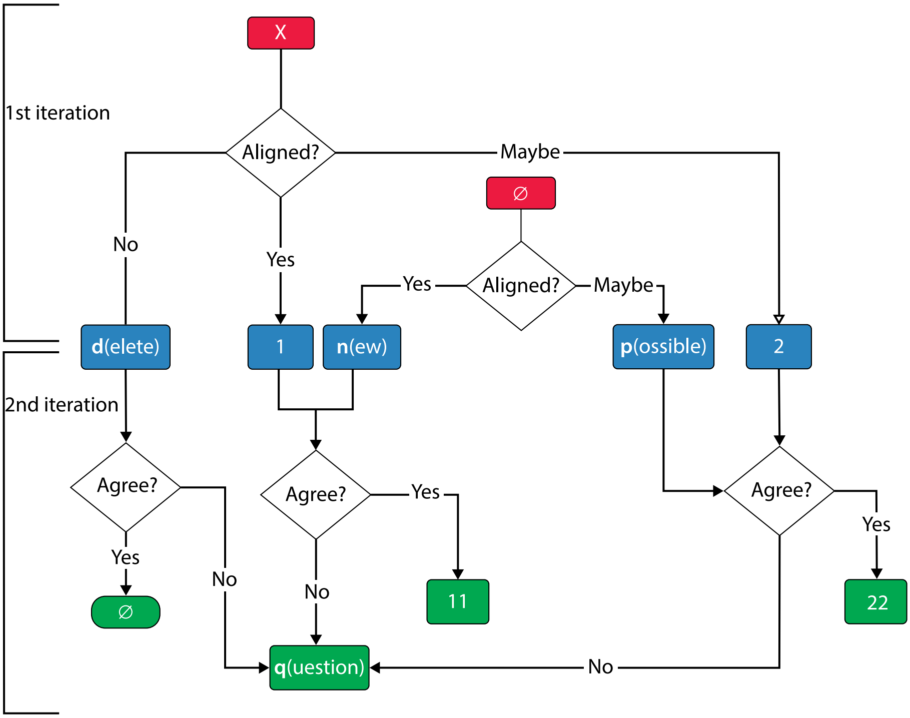


Figure 2. Align flow for peer review alignment

them was reached. Some cases became precedents for principles and exceptions, which provided a good base for the gold rules to be used by other linguists and bi-/multilingual datasets builders.

2.2 Alignment manifesto

To make the alignment as unambiguous as possible, we first developed the high-level principles, or alignment manifesto, and then implemented them in the form of more definite alignment rules.

Principle 1 stipulates that alignment should be based on word representation in the languages and not on the context. In other words, only tokens with clear semantic correspondences in both languages should be aligned, while those added or omitted due to context necessity, literal purposes, or to prevent tautology in a language should not. Example (1) shows that the Chinese expression 他们两人 does not have its counterpart in the Russian language but is embedded in the word forms of the Russian words making it clear that a speaker is talking about them, which is the reason for omitting them in a Russian context that does not tolerate tautology.

(1)	**they**	**two person**	indeed	"string of pearl and jade"[3]
	tamen	**liangren**	zhenshi	zhulianbihe
	他们	**两人**	真是	珠联璧合
			поистине	прекрасная пара
			poinstine	prekrasnaya para
			indeed	lovely/nice couple

'(They) are indeed a lovely couple'

Principle 2 is adopted from previous research and postulates that there are S(ure) and P(ossible) types of alignment. As it is shown in (2), Chinese 庆祝 that literally translates as "to celebrate" is expressed in Russian as "in honor of," which is a possible but not a direct translation of the Chinese token. Therefore, following the "paradigmatic, not syntagmatic" principle, 庆祝 is aligned as a P-alignment with the Russian "в честь."

(2)	10 month 1 day	FUT	run	**celebrate**	China	establish	70_years
	10 yue 1 ri	jiang	juban	**qingzhu**	zhongguo	chengli	70_zhounian
	10 月 1日	将	举办	**庆祝**	中国	成立	70周年
	milirary_parade						
	yuebingshi						
	阅兵式						

Китай	проведет	военный	парад	**в честь**	70-летия	КНР
Kitai	provedet	voennyi	parad	**v chest'**	70-letiya	KNR
China	organise-FUT	military	parade	**in honor**	70-year-PL	China

'(On 1st October) a military parade will be held to celebrate the 70th anniversary of the establishment of China'.

Figure 3 explains the difference between the so-called paradigmatic and syntagmatic alignments. The concepts of paradigmatic and syntagmatic relations come from Saussure's Basic Principles of Structural Linguistics, where he calls a relation between different grammatical roles of words in a sentence syntagmatic; and a relation between two interchangeable words paradigmatic. Therefore, in our project, words/expressions are aligned as S-alignments when they are the direct/evident correspondences of each other, as in the example in the center, where the Chinese and Russian words "beer" in bold represent the S-alignment. The left example on the diagram is an example of "paradigmatic" alignment, where a source word aligned to a rare synonym in the target language. Italics indicate that this is a P-alignment, as in the Russian language, "beer" is translated euphemistically, with a possible but not a straightforward word/expression. An example of

[3] idiom. perfect pair

"syntagmatic" alignment is on the right. Chinese 一杯 corresponds to the ending of the Russian word "пива" that shows that it is only one glass of beer. It seems more complex and does not suit the purposes of building a gold standard of rules. The gold standard should be able to train the algorithm for finding sure/possible alignments for a language pair without any grammatical distinction – just with the ability to classify corresponding translations according to their commonness in respect to a searched word in a source language.

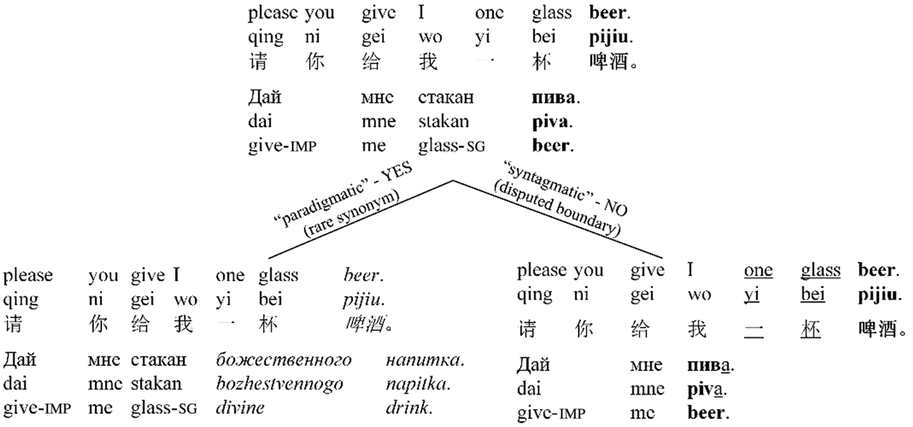

please you give I one glass beer.
qing ni gei wo yi bei pijiu.
请 你 给 我 一 杯 啤酒。
Дай мне стакан пива.
dai mne stakan piva.
give-IMP me glass-SG beer.
"paradigmatic" - YES (rare synonym)
"syntagmatic" - NO (disputed boundary)
please you give I one glass beer.
qing ni gei wo yi bei pijiu.
请 你 给 我 一 杯 啤酒。
Дай мне стакан божественного напитка.
dai mne stakan bozhestvennogo napitka.
give-IMP me glass-SG divine drink.
please you give I one glass beer.
qing ni gei wo yi bei pijiu.
请 你 给 我 一 杯 啤酒。
Дай мне пива.
dai mne piva.
give-IMP me beer.

Figure 3. "Paradigmatic" vs. "Syntagmatic" alignments

2.3 Alignment rules

2.3.1 *Punctuation*

Punctuation is an essential part of the written form of every language. With proper punctuation marks, people differentiate ideas on paper when a speaker's voice is not heard. Thus, in a written context, punctuation marks are as crucial as words themselves.

Besides auxiliary verbs, question/exclamation words/expressions (like "What a nice ..." in English, "真 ... *(好看)*" in Chinese, or "Какой/какая ..." in Russian), and punctuation marks, which are the most expressive markers in the written discourse, Chinese abounds in modal particles that do not find proper word-correspondences in Russian. Particles express emotions, sometimes perform an exclamation or interrogative function, and could be divided into functional and non-functional particles. When a modal particle performs an interrogative/exclamation function, it is S-aligned with a punctuation mark in a Russian sentence, see (3). When a particle is non-functional and serves as an emotional decoder in written discourse, it is marked as a P-alignment to a punctuation mark in

Russian. Sample (4) is an example of a non-functional particle alignment. Bold tokens stand for S-alignment and underscored tokens represent P-alignment, that is question marks in both languages, and a bold and underscored token is an S-alignment for a bold token in another language and a P-alignment for a respective underscored token, that is Chinese 呀 is a P-alignment to a question mark in the Russian sentence.

(3) You be this way think **Q** **?**
Nin shi zhe yang renwei **ma** **?**
您 是 这 样 认为 **吗** ?
Вы думаете **?**
Vy dumaete **?**
You think **?**
'Do you think so?'

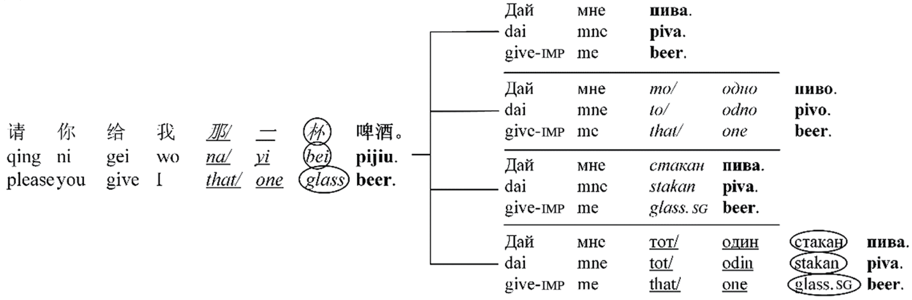
(4) Search who EMPH **?**
Zhao shui ya **?**
找 谁 呀 **?**
Кого надо **?**
Kogo nado **?**
Who.ACC need **?**
'Who are you looking for?'

Like in (3), the modal particle 吗 performs an interrogative function: even without seeing a sentence, an interlocutor can understand that it is a question because it is the "duty" of 吗 here to set a questioning tone; therefore, 吗 together with a question mark in Chinese is a sure alignment to a question mark "?" in Russian, as in the latter a question is expressed by means of intonation primarily.

2.3.2 *Pronouns and classifiers*

The Chinese language is well-known for its richness in classifiers. Russian, on the other hand, does not have that many classifiers; in rare cases, when it does, it often considers their use redundant or excessive. This results in a lack of correspondence with Chinese classifiers. In addition, the flexibility (that is the same sentence could be expressed/translated with or without pronouns without losing meaning) of the Russian language shown in (5) makes the alignment issue sharper. Following Principle 1, we set an additional rule for classifiers. It postulates that a Chinese classifier is left without alignment when a classifier or a similar word is omitted in Russian. In Russian, a demonstrative pronoun, a numeral or a so-called "classifier," that is the word denoting a container as in the case of (5),[4] is aligned with the Chinese classifier structure as an S-alignment. When a demonstrative and a "container"-classifier appear in Russian translation, every word is aligned as an S-alignment separately.

(5)


请 你 给 我 那/ 一 杯 啤酒。
qing ni gei wo na/ yi bei pijiu.
please you give I that/ one glass beer.
Дай мне пива.
dai mne piva.
give-IMP me beer.
Дай мне то/ одно пиво.
dai mne to/ odno pivo.
give-IMP me that/ one beer.
Дай мне стакан пива.
dai mne stakan piva.
give-IMP me glass.SG beer.
Дай мне тот/ один стакан пива.
dai mne tot/ odin stakan piva.
give-IMP me that/ one glass.SG beer.

'Give me (that/ one/ glass/ one glass of) beer.'

2.3.3 *Chinese particles and verb complements*

Compared to inflected or fusional Russian, Chinese, as a representative of isolating languages, has many particles, some of which perform as members of nominal, verbal, and adverbial structures, and some represent the tense category. In the Russian language, these particles find their representation in inflectional changes, therefore, are attributed to "words on the left" in Chinese and aligned with the proper correspondences in Russian. For example, 的 is aligned together with the corresponding pronoun, noun, or adjective to a proper Russian token, 得 is aligned with a Russian verb together with the main verb in Chinese, and 地 is attributed to the corresponding adjective in Chinese. The same rule is applied to Chinese tense particles 了, 着, 过, etc.

Chinese directional compounds 去, 来, 下来, 下去, 上来, 上去 and resultative compounds 到, 见, 在一起, etc. are aligned together with the preceding Chinese main verbs (see (6) to (8)). However, when Chinese 去, 来, 到 are used not as compounds but as independent members of a sentence, that is as verbs or prepositions, they are aligned with their respective correspondences in Russian sentences, as in (9).

[4] Here, bold tokens stand for S-alignments, italicized represent P-alignments, underlined and circled show S-alignments and draw a clear line which tokens are aligned as S-alignments with each other.

(6) He call PST one CLF have-meat POSS cabbage_soup, immediately **sit**
ta jiao le yi fen dairou de baicaitang, jiu **zuo**
他 叫 了 一 份 带肉 的 白菜汤, 就 **坐**
down_DIR_COMP.
xialai.
下来.
Он спросил себе щей с мясом и **сел.**
on sprosil sebe shchei s myasom I **sel.**
he ask-PST himself shchi with meat and **sit-PST.**
‘He asked for cabbage soup and sat down.’

(7) he look PROG Katerina, just **think RES_COMP** …
ta qiao zhe kajielinna, zhishi **xiang dao** …
他 瞧 着 卡捷琳娜, 只是 **想 到** ……
Глядя на неё, он только **подумал** …
glyadya na nee, on tol’ko **podumal** …
look_PTCP at she-POSS, he just **think-PST** …
‘Looking at her (Katerina), he just thought …’

(8) she POSS pupil open V-AUX so big, almost and black POSS
ta de tongkong zheng de zheme da, jihu he heise de
她 的 瞳孔 睁 得 这么 大, 几乎 和黑色 的
iris_circle **join together PST.**
hongmojuan **he zaiyiqi**
虹膜圈 **合 在一起了.**
Зрачки её были расширены так, что почти **сошлись** с
zrachki ee bili rashireny tak, chto pochti **soshlis’** s
pipul.PL she-POSS be-PST dilated-PART so, that almost **get_together-PST** with
чёрными ободками радужки.
chernymi obodkami raduzhki.
black.PL rim.PL iris-GEN.
‘Her pupils opened so wide that they were almost the size of the black circles of her iris.’

(9) he **go-PST** to Moscow.
on **poehal** v Moskvu.
он поехал в Москву.

她 到 莫斯科 **去 了。**
ta dao mosike **qu le.**
he to Moscow **go PST.**

她 **去** 莫斯科 **了。**
ta **qu** mosike **le.**
he **go** Moscow **PST.**

‘He went to Moscow.’

2.3.4 *Prepositions*

In comparison to (5), Example (9) shows Chinese flexibility. The Chinese language allows, with a verb position change, for some prepositions to be omitted in certain sentences without any change in meaning. In contrast, in Russian, some verbs are not used without prepositions, such as the verb "to go (somewhere or to a certain destination)." Therefore, when Chinese does not have a verbal preposition, low right example in (9), and a Russian verb is used mainly with a preposition in a designated context, the Chinese verb is aligned with the Russian verb and its preposition, but, as it is shown in (9), as the S-alignment with the verb "поехал" and as the P-alignment with the preposition "в".

Some peculiar preposition usages are also found in Chinese. For instance, Chinese 在, when used as a preposition, often forms a frame structure, 在… 上 (on… surface), whereas in Russian, it is expressed by one preposition only, see (10). Thus, to mitigate the grammatical differences between languages, the whole Chinese prepositional frame structure, 在 and 上, is aligned with the Russian preposition "на".

(10) Alexandra and Ana together sit **on** one CLF small sofa **surface**
yalishandela he ana yiqi zuo **zai** yi zhang xiao shafa **shang.**
亚历山德拉 和 安娜 一起 坐 **在** 一 张 小 沙发 **上**。
Александра и Анна сели вместе **на** маленьком диване.
Aleksandra i Anna seli vmeste **na** malen'kom divane.
Aleksandra and Anna sit-PST together **on** small sofa.
'Aleksandra and Anna sat together on a small sofa.'

Discrepant prepositional structures deserve special attention among all prepositional rules. Chinese and Russian not only differ in the lack or presence of prepositions but also have so-called "mismatched prepositional structures." At first sight, it looks that the Chinese structure 从这个案子脱身 (from this case free oneself) from (11) is easily aligned with the Russian "*сбросить с себя этот груз*" (throw off from oneself this burden). Moreover, all the components seem common between Chinese and Russian: 这个案子 (this case) corresponds to "*этот груз*" (this burden), 脱 (free) to "*сбросить*" (throw off), 身 (oneself) to "*себя*" (oneself), 从 (from) to "*с*" (from). However, a closer analysis shows that the Chinese 从 (from) refers to "(to get rid of) this case," whereas the Russian preposition "*с*" (from) is attributed to "(throw off from) oneself, from your body," if the Russian sentence is translated literally. Therefore, in the Chinese structure 从……脱身 (from … free oneself), 脱身 (free oneself) should be aligned as an S-alignment with the entire Russian expression "*сбросить с себя*" (throw off from oneself) and the Chinese preposition 从 (from) is to be left without a pair as in Russian "*этот груз*" (this burden) does not require any prepositions.

(11) Lao Xing have_to from this CLF case **free_oneself.**
lao Xing zhihao cong the ge anzi **tuoshen.**
老 邢 只好 从 这 个 案子 **脱身**。
Итак, ему ничего не оставалось, как **сбросить с себя** этот
Itak, emu nichego ne ostavalos', kak **sbrosit' s sebya** etot
So, he-DAT nothing no leave-PST, but **throw_off from oneself** this
груз.
gruz.
burden.
'(Lao Xing) he had nothing to do but to get rid of this burden.'

2.3.5 *Chinese verbs "to be" and "to have"*

The fact that Russian and Chinese belong to different language families causes more issues to be elaborated separately. Russian is considered to be a "to be"-language, and Chinese can be classified as a "to have"-language (see Freeze 1992). Russian flexibly uses the verbs "to be" and "to have," whereas Chinese, together with lingua franca English, has stricter rules and more regular usage. Therefore, following Principle 1, the verbs "to be" and "to have" are aligned when they have equivalents in Russian. For instance, (12) shows that in the present tense the verb "to be" is not translated/used in Russian, but in the past, as in (13), it appears and is aligned with the Chinese 是 (to be).

(12) I **be** student.
wo **shi** xuesheng.
我 **是** 学生。
Я студент.
ya student.
I student.
'I am a student.'

(13) You why with he break_up PST? he **be** CLF idiot.
ni weishenme gen ta fenshou le? ta **shi** ge bendan.
你 为什么 跟 他 分手 了？ 他 **是** 个 笨蛋。
Почему ты с ним рассаталась? Он **был** дурак.
pochemu ty s nim rasstalas'? on **byl** durak.
Why you with he-ABL break_up-PST? he **be-PST** idiot.
'Why did you break up with him? He was an idiot.'

The verb "to have" does not depend on the tense (see (14)). Still, it sometimes could be omitted in the Russian language, as in (15), sometimes is embedded in negation or adverbs, as in (16), and even sometimes represented by synonyms (see (17)), which can only be P-alignments. Due to the word limits, we have presented different examples for adverb cohesion and synonym replacement. As for negation, it is usually presented as an S-alignment of Chinese 没有 to Russian "*нет*" in a present form.

(14) he **have/ FUT_have** house.
ta **you/ jiangyou** jia.
他 **有/ 将有** 家。
У него **есть/ будет** дом.
u nego **est'/ budet** dom.
At he-GEN **have/ have-FUT** house.
'He has/will have a house.'

(15) already **have** good many person EMPH.
yijing **you** hao duo ren la.
已经 **有** 好 多 人 啦。
Уже очень много.
uzhe ochen' mnogo.
already very many.
'There are already many people.'

(16) factory_building **in_total have** forty_eight CLF window.
changfang **yigong you** sishiba shan chuanghu.
厂房 **一共 有** 四十八 扇 窗户.
Вэтомпомещении **всего** сорок восемь окон.
v etom pomeshenii **vsego** sorok vosem' okon.
in this building **in_total** forty eight window.PL.
'There are forty-eight windows in this (factory) building in total.'

(17) small group nowadays just *have* organiser one person.
xiao zu muqian jin *you* zuzhizhe yi ren.
小 组 目前 仅 有 组织者 一 人。
Кружок пока что *состоял* только из одного организатора.
kruzhok poka_chto *sostoyal* tol'ko iz odnogo organizatora.
Club as_for_now *consist-PST* just of one-GEN organiser.
'The was only one person in the club; it was an organiser.'

2.3.6 *Alignment of speech figures*

The last rule we would like to present is the alignment of the so-called "translation peculiarities". What is evident in one language can be difficult to understand in another; thus, translators use communicative or adaptational translation methods to "rewrite" a source text so that it conforms to the rules of a target language and is smoothly acceptable by a target language. In (18), which is the translation of the Russian text into Chinese, a metonymy is included, that is "clothes" is used to denote a “person.” However, this metonymy may interfere with the smooth understanding once translated into Chinese, so the metonymy was opened and translated into “the widow that wore a starched underskirt.” Even though we are aware of this language phenomenon, following Principle 1, we aligned the Russian adjective “starched” to the Chinese noun “starch” and attributed to it possessive particle “de,” leaving another part of metonymy, 衬裙上过 (put on the underskirt), unaligned. By doing so, we both follow our main Principle 1 and leave the ground for learners and researchers to study how speech figures are presented in the two respective languages.

(18) disappear-PRF-PTCP.SG, that[…] **starched** widow-DIM not have_time …
ischez, chto[…] **krahmal’naya** vdovushka ne uspela …
исчез, что[…] **крахмальная** вдовушка не успела …
没 了 踪影, 衬裙 上过 **浆** **的** 寡妇 根本
mei le zongying, chenqun shangguo **jiang** **de** guafu genben
not PST trace, underskirt put_on **starch** **POSS** widow simply
来不及 …
laibuji …
do_not_have_time …
‘(he) disappeared (so fast) that the widow in a starched dress did not have time to …’

Above are the main rules of our gold dataset. Due to space limitations and since the scope of the paper is to describe some general guidelines for building a Russian-Chinese gold dataset and test its role in algorithm training, some straightforward rules or those similar to ones already mentioned – like the alignment of Chinese auxiliary particles 被 and 把 that are similar in alignment principle with compounds (see Section 3.3.3) – have been omitted.

3. Evaluation

Among several existing machine learning models for word alignment (Dou & Neibig 2021; Nagata, Chousa & Nishino 2020; Li et al. 2019), Awesome Align by Dou and Neibig was chosen due to the availability of the code, ease in use, and, importantly, due to its sound performance on English-Chinese language pair (13.6% AER vs. 36.5% by Li et al. 2019). All the papers mentioned above use BERT as their base model; however, they perform worse or comparably to the chosen Awesome Align.

For a baseline, we used two statistical models: IBM Model 1, which is more straightforward, and fast-align, which considers the words outside of any context and is a reparametrization of IBM Model 2.

Due to significant language differences, we applied several tokenization options for statistical algorithms: BPE (byte-pair encoding) (Gage 1994) and MyStem (Segalovich 2003). MyStem is an algorithm that provides word lemmatization, that is bringing a word to its "normal" form (better – > good, walking – > walk, etc.). We used MyStem lemmatization and BPE for Russian sentences; and single character tokenization for Chinese. In pursuit of better results, we upgraded MyStem with lemmatization, which resulted in better model performance (see Table 1).

Table 1. Statistical models' results

Method	Training	Ru-Zh AER
IBM-model 1	No data preparation	75.2%
	Lemmatized data	67.8%
	BPE-tokenized data	79.5%
Fast Align	No data preparation	69.3%
	Lemmatized data	**61.1%**
	BPE-tokenized data	71.0%

All models were trained on 3.5 million (all available at that time) words from RuZhCorp and underwent training on three datasets: no preprocessing for either language; MyStem lemmatization for Russian, no preprocessing for Chinese; and BPE tokenization for both languages. Table 1 shows that fast-align with lemmatization had significantly better performance and thus was chosen as the primary baseline for our training model.

Awesome Align uses vectorized word representations of a pretrained multilanguage model. Each token (word, character, or punctuation sign) is represented with an ordered set of numbers, called a vector or embedding that also depends on the context. Such embeddings can be learned from multilingual but not aligned corpora, which allows for the use of big publicly available datasets. The model then calculates the distance between each pair of Russian and Chinese tokens; when that distance is lower than a predefined threshold, the pair is considered aligned.

Dou and Neibig (2021) proposed several ways to fine-tune the model. In a sentence-aligned parallel corpus, 15% of randomly chosen tokens in both languages are replaced with a unique [MASK] token, with a random token or left unchanged, with the probabilities of 80%, 10%, and 10%, respectively. Given a pair of masked sentences in two languages, the model learns to reconstruct an original token by itself.

Awesome Align, which can use different models as its core algorithm, also allows the model to be fine-tuned using a word-level pre-aligned bilingual parallel corpus. A combination of translation language modeling, self-training objective (a method similar to EM algorithm), and other methods described in the Awesome Align paper was the objective for such fine-tuning. Two algorithms were used for the experiments:

1. MultiBERT (also used by Dou and Neibig (2021)) was further trained on a multilingual corpus of Wikipedia articles. The model learned the embeddings using the masking approach described above and predicted whether a given sentence follows another sentence in a text.
2. LaBSE is trained by the same masking approach and by a translation model with parallel corpora. It learned to predict whether two sentences in two languages are aligned in a parallel corpus.

We have evaluated word alignments that can be extracted directly from publicly available versions of MultiBERT and LaBSE. Further, we additionally trained those models on RuZhCorp (~700,000 sentence pairs). As a final step, we fine-tuned those models on gold set of data (over 350 sentence pairs) manually annotated by humans. To compare the quality of the models, we use the AER metric introduced by Och and Ney (2000). All RuZhCorp texts (3.5 million words) were used for training. When training on annotated sentences from the novel dataset, AER is calculated on the rest of the dataset (test set of 102 sentence pairs).

In total, we conducted eight experiments, the results of which are shown in Table 2.

Table 2. AER of MultiBERT and LaBSE models after different trainings

Method	Training	Ru-Zh AER
Baseline (statistical model)	unsupervised on the parallel corpus	61.1%
AA – MultiBERT	Bare model (not explicitly trained for word alignment)	38.2%
	Pre-trained by Dou and Neibig	31.5%
	Fine-tuned on RuZhCorp (5 epochs)	28.7%
	Finer-tuned on the gold dataset	28.3%
AA – LaBSE	Bare model	31.8%
	Fine-tuned on RuZhCorp (1 epoch)	19.8%
	Finer-tuned on the gold dataset	**18.9%**

Table 2 shows that LaBSE achieves the best AER of 18.9% and MultiBERT follows with 28.3% only. That shows two facts: first, a gold dataset may improve the algorithms' performance (by 0.9 percentage points in the LaBSE case); second, in either fine-tuning scenario, LaBSE performance exceeds that of MultiBERT.

Summing up, in the absence of previous work on Russian-Chinese word alignment, we have compared our results with other resemblant pairs of non-similar languages that include one European and one East-Asian language: Li et al. (2019) list 36.57% as their best result for Chinese-English, Dou and Neibig (2021) show an AER of 37.4% for Japanese-English while providing a much lower 13.9% AER for Chinese-English. Therefore, our results are in line with or better than the previous research on similar language pairs and may become a valuable benchmark for future research on Russian-Chinese word alignment.

4. Conclusion

Word-alignment is a relatively novel and complicated task. In this paper, we described how we built a gold dataset of Russian-Chinese word-aligned sentences and the role of this dataset in algorithm training. We established the manual alignment guidelines for the Russian-Chinese language pair and showed that simple spreadsheets are helpful in the construction of a gold dataset as they allow for many-to-many alignments and peer-review. Having evaluated different models, we found that LaBSE with fine-tuning showed better results, and so we applied it to the RuZhCorp existing dataset. The alignment results are available online at < https://linghub.ru/rnc_parallel_chinese/search >

The good results after fine-tuning on the gold dataset are promising for further work, which we expect to undertake. First, we plan to increase the manually aligned dataset and train the model on a more extensive training set. Second, we hope to experiment with similar algorithms on other parallel corpora of the Russian National Corpus. Third, we want to try newer core models of the BERT family. Fourth, we plan to implement a translation relevance mechanism based on the word alignment, that is a mechanism that differentiates between more and less likely translations and arranges the sentences from sure correspondences to context or P-translations. We hope that our work, which is on par with current state-of-the-art models for similar language pairs, could facilitate the development of new word alignment methods and that the results of Russian-Chinese corpus alignment can benefit both students and professionals.

Funding

This work was supported by the “Development of online reference tools based on RuZhCorp to create an interactive educational environment” project funding from the Students’ Academic Development Center of Higher School of Economics, Russia, and Jiangsu Province Excellent Postdoctoral Researcher Funding (江苏省卓越博士后计划), PRC.

Acknowledgements

We thank Kirill Semenov for his kind assistance during the project and for the allocated time for fruitful discussions. We also thank the reviewers for their valuable comments and everyone who helped us to proof the manuscript.

Abbreviations

ABL	ablative case	PL	plural
CLF	classifier	POSS	possessive
DAT	dative case	PRF	perfect
DIM	diminutive	PST	past
DIR_COMP	directional compound	PTCP	participle
EMPH	emphatic marker	RES_COMP	resultative compound
FUT	future	SG	singular
GEN	genitive case	V-AUX	auxiliary verb
IMP	imperative		

References

Alkhouli, Tamer & Ney, Hermann. 2017. Biasing attention-based recurrent neural networks using external alignment information. In Proceedings of the Second Conference on Machine Translation, Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann et al. (eds), 108–117. Copenhagen: Association for Computational Linguistics.

Bahdanau, Dzmitry, Cho, Kyunghyun & Bengio, Yoshua. 2014. Neural machine translation by jointly learning to align and translate. <https://arxiv.org/abs/1409.0473> (30 July 2022).

Brown, Peter F., Cocke, John, Pietra, Stephen Della, Pietra, Vincent J. Della, Jelinek, Frederick, Lafferty, John D., Mercer, Robert L. & Roossin, Paul S. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2): 79–85.

Brown, Peter F., Pietra, Stephen A., Pietra, Vincent J. Della & Mercer, Robert L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2): 263–311.

Chen, Jiang & Nie, Jian-Yun. 2000. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In Proceedings of the Sixth Conference on Applied Natural Language Processing, 21–28. Seattle WA: Association for Computational Linguistics.

Chen, Wenhu, Matusov, Evgeny, Khadivi, Shahram & Peter, Jan-Thorsten. 2016. Guided alignment training for topic-aware neural machine translation. <https://arxiv.org/abs/1607.01628> (30 July 2022).

Cysouw, Michael & Wälchli, Bernhard. 2007. Parallel texts: Using translational equivalents in linguistic typology. *Language Typology and Universals* 60(2): 95–99.

Davis, Mark & Dunning Ted E. 1995. Query translation using evolutionary programming for multi-lingual information retrieval. In *Query Translation Using Evolutionary Programming for Multi-lingual Information Retrieval*, John R. McDonnell, Robert G. Reynolds & David B. Fogel (eds), 175–185. Cambridge MA: The MIT Press.

Dempster, Arthur P., Laird, Nan M. & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1): 1–38.

Dou, Zi-Yi & Neubig, Graham. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Paola Merlo, Jorg Tiedemann, Reut Tsarfaty (eds), 2112–2128. Stroudsburg PA: Association for Computational Linguistics.

Freeze, Ray. 1992. Existentials and other locatives. *Language* 68: 553–595.

Gage, Philip. 1994. A new algorithm for data compression. *The C Users Journal archive* 12: 23–38.

Graça, João V., Pardal, Joana Paulo, Coheur, Luisa & Caseiro, Diamantino Antonio. 2008. Building a golden collection of parallel multi-language word alignment. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 986–993. Marrakech: LREC.

Han, Nway Nway & Thida, Aye. 2019. Annotated guidelines and building reference corpus for Myanmar-English word alignment. *International Journal on Natural Language Computing* 8(4): 25–38.

Kruijff-Korbayová, Ivana, Chvátalová, Klára & Postolache, Oana. 2006. Annotation guidelines for Czech-English word alignment. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard et al. (eds), 1256–1261. Genoa: ELRA.

Kuhn, Jonas. 2004. Experiments in parallel-text based grammar induction. *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 470–477. Barcelona: Association for Computational Linguistics.

Lambert, Patrik, Gispert, Adria, Banchs, Rafael & Mariño, Jose B. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation* 39(4): 267–285.

Li, Jinji, Kim, Dong-Il & Lee, Jong-Hyeok. 2008. Annotation guidelines for Chinese-Korean word alignment. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (Eds), 518–524. Marrakech: ELRA.

Li, Xintong, Li, Guanlin, Liu, Lemao, Meng, Max & Shi, Shuming. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Preslav Nakov & Alexis Palmer (eds), 1293–1303. Florence: Association for Computational Linguistics.

Macken, Lieve. 2010. An annotation scheme and Gold Standard for Dutch-English word alignment. Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 10), Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani et al. (eds), 3369–3374. Valletta: ELRA.

Mayer, Thomas & Cysouw, Michael. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokić, Michael Cysouw (eds), 54–62). Avignon: Association for Computational Linguistics.

Nagata, Masaaki, Chousa, Katsuki & Nishino, Masaaki. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing [online], Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds), 555–565. Stroudsburg PA: Association for Computational Linguistics.

Nie, Jian-Yun, Simard, Michel, Isabelle, Pierre & Durand, Richard. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 74–81. New York NY: Association for Computing Machinery.

Och, Franz Josef & Ney, Hermann. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 440–447. Hong Kong: Association for Computational Linguistics.

Och, Franz Josef & Ney, Hermann. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–51.

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In *The Lexical Typology of Semantic Shifts*, Päivi Juvonen & Maria Koptjevskaja-Tamm (eds), 157–176. Berlin: De Gruyter Mouton.

Sahlgren, Magnus & Karlgren, Jussi. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering* 11(3): 327–341.

Segalovich, Ilya. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, Hamid R. Arabnia & Elena B. Kozerenko (eds), 273–280. Las Vegas NV: CSREA Press.

Stahlberg, Felix, Saunders, Danielle & Byrne, Bill. 2018. An operation sequence model for explainable neural machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Tal Linzen, Grzegorz Chrupała & Afra Alishahi (eds), 175–186. Brussels: Association for Computational Linguistics.

Stengel-Eskin, Elias, Su, Tzu-Ray, Post, Matt & Van Durme, Benjamin. 2019. A discriminative neural model for cross-lingual word alignment. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Sebastian Padó & Ruihong Huang (eds), 910–920. Hong Kong: Association for Computational Linguistics.

Wälchli, Bernhard & Cysouw, Michael. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3): 671–710.

Yadav, R. K. & Gupta, Deepa. 2010. Annotation guidelines for Hindi-English word alignment. Proceedings of the International Conference on Asian Language Processing IEEE, 293–296. Harbin: IALP.

Author Query

- Please provide citation indicator for missing this reference 'Bahdanau, Dzmitry, Cho, Kyunghyun & Bengio, Yoshua. 2014'.