Цзинь Ифан,
аспирант Высшей школы перевода (факультета)
МГУ имени М.В. Ломоносова; e-mail: 1554583267@qq.com

РАЗВИТИЕ МЕЖЪЯЗЫКОВЫХ БОЛЬШИХ ДАННЫХ И КОРПУСА КИТАЙСКОГО ЯЗЫКА

В данной статье рассматриваются межъязыковые большие данные, которые в последние годы привлекли внимание китайских и российских исследователей. Исследуются платформа анализа межъязыковых больших данных YeeSight и создание корпуса китайского языка. Цель статьи — рассмотреть преимущества и недостатки платформы YeeSight и продемонстрировать развитие корпуса китайского языка в последние годы. Основное внимание уделено выяснению и описанию основного набора тэгов частеречной разметки корпусов китайского языка.

Ключевые слова: межъязыковые большие данные, платформа YeeSight, создание корпуса китайского языка.

1. YeeSight на фоне «больших данных»

Термин «большие данные» связывают с Клиффордом Линчем, редактором журнала "Nature", этот термин был первый предложен по аналогии с расхожими в деловой англоязычной среде метафорами «большая нефть», «большая руда» в 2008 году. Начиная с 2012 года, большие данные постепенно стали одной из самых горячих тем. «Большие данные» — это масса новых задач, касающихся общественной безопасности, глобальных экономических моделей, неприкосновенности частной жизни, устоявшихся моральных правил, правовых отношений человека, бизнеса и государства [Майер-Шенбергер, Кеннет Кукьер, 2014: 8]. Дуг Лэни, аналитик "Gartner", считает, что у больших данных три главных признака — по-английски это три V [Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung, 2014: 4]: Volume — большой объём данных, Velocity — скорость поступления (данных не просто много, а их становится всё больше и больше), Variety — данные разнообразны (есть структурированные, а есть плохо структурированные, с которыми надо работать параллельно [Садовничий, 2017: 12].

Европейские и американские компании по анализу больших данных дали свои ответы. Компания Google использовала технологию анализа данных и предоставляла клиентам рекламы [Levy, 2012: 51].

В 2016 году Компания Cambridge Analytica анализировала данные клиентов Facebook, чтобы помочь Дональду Трампу выиграть президентские выборы в США. Американская разведывательная служба ЦРУ указала, что 90% информации мировой военной разведки можно получить из больших данных [Viktor Mayer-Schönberger, 2012: 33]. Анализ данных играл важную роль в войне между США и Бен Ладеном.

Кроме США и другие страны тоже разрабатывают технологию анализа больших данных. В 2016 году китайская компания GTCOM опубликовала платформу анализа межъязыковых больших данных — YeeSight.

Что такое межъязыковые большие данные? В октябре 2015 года компания GTCOM впервые предложила этот термин, который является новым фактором производства. Большие данные сегодня — это не просто научный термин. Это — глобальный феномен, фактор окружающей среды [Садовничий, 2017: 5]. Межъязыковые большие данные объединяет AI, языки, перевод и большие данные, применяются для стимулирования роста в широком спектре различных сфер, включают финансы, туризм, новые средства массовой информации и точный маркетинг. YeeSight — это платформа анализа межъязыковых больших данных, объединяет преимущества мощного машинного перевода и искусственного интеллекта, может помочь людям устранить языковые барьеры и расширить возможности приобретения знаний [Ю Янь, 2014: 14]. Пользователи могут искать информацию на своём родном языке и получать все соответствующие результаты на китайском, английском, русском, немецком и любом другом языке.

Достоинство платформы YeeSight заключается в том, что практические проблемы она умеет переводит в статистические, при этом для решения данных проблем используется система машинного перевода (система YeeCloud). Поэтому суть системы YeeSight заключается в том, чтобы использовать аналитические модели больших массивов данных в помощь нуждающимся в межъязыковой информации клиентам (объём межъязыковой информации больше, чем объём одноязычной информации) и переводить её на собственный язык клиента при посредстве машинного перевода.

Несмотря на то, что платформа YeeSight развивается довольно быстрыми темпами, она также сталкивается с многочисленными проблемами. Проблема в основном заключается в её системе перевода (система YeeCloud): во-первых, при переводе с разных языков эффективность перевода отличается. Например, перевод фразы с китайского на английский разительно отличается от перевода этой

же фразы с китайского на русский. Во-вторых, результаты перевода нестабильны и неравномерны, качество перевода некоторых текстов оставляет желать лучшего. Перевод — это перевыражение. Если всякое речевое произведение представляет собой в известном смысле материальное оформление отражения фрагмента действительности сознанием индивида, то перевод является отражением отражения [Гарбовский, 2007: 10]. В-третьих, качество перевода зависит от тематики текста, например, новости спорта сложнее для перевода, чем политические новости.

Успех платформы YeeSight заключается в том, что она объединяет преимущества мощного машинного перевода и искусственного интеллекта, расширяет сферу исследований перевода, содействует развитию теории перевода на практике. Недостатком же является то, что в эпоху «больших данных» размер данных относителен, и при решении каких-то более сложных задач естественно всегда требовались и требуются большие объёмы данных. Например, на данный момент объём данных ещё недостаточно велик для решения проблемы повышения точности перевода с китайского языка на русский, для этого YeeSight придётся развивать аналитические модели больших массивов данных.

Способ обработки задач в YeeSight — классический пример использования больших данных, который имеет огромное прикладное значение для решения практических проблем этой среды.

2. Создание корпуса китайского языка

Платформа анализа межъязыковых больших данных YeeSight, как и любая языковая платформа, много значит для создания национального корпуса китайского языка.

Что такое национальный корпус китайского языка? Ещё несколько десятилетий назад работа китайских лингвистических исследований выполнялась вручную, это требовало очень много времени. Философы, психологи и другие специалисты отмечают, что в будущем социально защищённым может считаться лишь тот человек, который способен гибко перестраивать направление и содержание своей деятельности в связи со сменой технологий или требований рынка [Зубов, Зубова, 2004: 9]. С развитием компьютерных технологий стало возможным проводить лингвистические исследования быстрее. Под лингвистическим корпусом понимается совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой [Перцов, 2006: 319]. Учёные могли создать корпус текстов на базе китайского языка, охватыва-

ющего от миллионов до десятков миллиардов лексических единиц, особенностью является использование больших объёмов текстовой информации, сведённой в единую базу, специальным образом размеченной и именуемой корпусом. Корпусная лингвистика — раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (корпусов текстов) с применением компьютерных технологий. Под лингвистическим, или языковым, корпусом текстов понимается большой, представленный в машиночитаемом виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач [Захаров, Богданова, 2011: 7].

Первым китайским компьютерным корпусом считается Научно-технический английский корпус, который был создан в 1980-е в Шанхайском транспортном университете. Потом корпусная лингвистика в Китае быстро развивается, существуют различные корпуса, такие как

древний китайский корпус
(http://corpus.zhonghuayuwen.org/ACindex.aspx),

современный китайский корпус
(http://corpus.zhonghuayuwen.org/CnCindex.aspx).

Для академических исследований некоторые китайские университеты построили свои собственные китайские корпусы, самым известным из них является китайский корпус «Пекинский университет» (http://ccl.pku.edu.cn/corpus.asp). В корпусе сохраняется не только большое количество китайской информации, но и сохраняется большая часть английской информации. Помимо университетов существуют также крупные компании, которые разрабатывают свои собственные корпусы, например,

Google Inc
(http://storage.googleapis.com/books/ngrams/books/datasetsv2.html),

GTCOM
(http://www.yeesight.com/),

Iflytek
(https://www.xfyun.cn/),

Sogou
(http://www.sogou.com/labs/dl/c.html).

Чтобы разработать единый стандарт, институт прикладной лингвистики при Министерстве образования КНР предложил «Принцип частеречной разметки в обработке современной ки-

тайской информации», устанавливающий конкретный стандарт морфологической разметки.

Основной набор тэгов частеречной разметки в корпусах китайского языка

Часть речи	Пример	Тэг
Имя существительное	熊猫 (панда)	n
Существительное, означающее азимутальное направление	西南 (юго-запад)	nd
Существительное времени	周四 (четверг)	nt
Существительное, указывающее местонахождение	窗前 (у окна)	nl
Имя собственное	习nhf	nh
Фамилия	近平nhg	nhf
Имя	(Си Цзиньпин)	nhg
Глагол	走 (идти)	v
Глагол направленного действия	过来 (прийти)	vd
Глагол, выполняющий функцию связки (специальное слово для связи подлежащего и сказуемого)	是 (являться)	Vl
Модальный глагол	能 (мочь)	vu
Географическое название	俄罗斯 (Россия)	ns
Названия учреждений, организаций и компаний	内务部 (Министерство внутренних дел)	ni
Прилагательное	聪明 (умный)	a
Числительное	一百 (сто)	m
Счётное слово	次 (раз)	q
Наречие	总是 (всегда)	d

Местоимение	我 (Я)	r
Союз	但是 (но)	c
Частица	是 (да)	u
Междометие	哎呦 (ой)	e
Звукоподражание	喵 (мяу)	o
Идиома, включая собственно фразеологию, пословицы и поговорки	多此一举 (В Тулу со своим самоваром ехать)	i
Аббревиатура	外交部 (Министерство иностранных дел)→(МИД)	j
Предлог	多亏了 (Благодаря)	p
Пунктуационный знак	w	w

Очевидно, что это принцип способствует созданию корпуса китайского языка. Вследствие этого человечеству в дальнейшем надо будет серьёзно задуматься над развитием национального корпуса.

Список литературы

Гарбовский Н.К. Теория перевода. М.: Издательство Московского университета, 2004. С. 10.

Перцов Н.В. О роли корпусов в лингвистических исследованиях // Труды международной конференции «Корпусная лингвистика–2006». СПб.: Изд-во С.-Петерб. ун-та, 2006. 319 с.

Садовничий В.А. Большие данные в современном мире. Доклад. М., 2017. 4 с.

Levy S. In the plex. How Google thinks works and shapes our lives M. Simon and Schuster, 2011. 51 p.

Mayer-Schönberger. Big Data: A Revolution Tat Will Transform How We Live, Work, and Tink Viktor, 2012. 33 p.

Min Chen, Shiwen Mao, Yin Zhang, Victor C.M. Leung. Big Data. Related Technologies, Challenges, and Future Prospects. Spinger. 2014. 4 p.

Yu Yang. Yeesight, Big Data Ecosystem-GTCOM, 2014. 14 p.

Jin Yifang,
Postgraduate Student at the Higher School of Translation and Interpretation, Lomonosov Moscow State University, Russia;
e-mail: 1554583267@qq.com

THE DEVELOPMENT OF INTERLINGUAL BIG DATA AND CORPORA OF THE CHINESE LANGUAGE

The article develops the idea that theoretical research work on interpretation in China began relatively late compared to the West. The Chinese theory of interpretation has been influenced by Western interpretation theories, among which the French translation school has had a particularly great influence on the Chinese theoretical foundations for interpretation didactics. The Chinese interpretation didactics has incorporated the most advanced Western theories of interpretation and is now moving toward scientification, systematization and professionalism, which has laid a solid foundation for professional training in the field of interpretation.

Key words: interlanguage big data, the YeeSight platform, creation of a corpus of the Chinese language.

References

Garbovskij N.K. Teorija perevoda [Theory of Translation] Moscow: Izdatel'st-vo Moskovskogo universiteta, 2007, pp. 10 (In Russian).

Levy S. In the plex. How Google thinks works and shapes our lives M. Simon and Schuster, 2011. 51 p.

Mayer-Schönberger. Big Data: A Revolution Tat Will Transform How We Live, Work, and Tink Viktor, 2012. 33 p.

Mayer-Schönberger, Kennet Kuk'er. Bol'shie dannye. Revolyuciya, kotoraya izmenit to, kak my zhivyom, rabotaem i myslim [Big Data. A Revolution That Will Transform How We Live, Work, and Think]. Per. s angl. Inny Gajdyuk. Moscow: Mann, Ivanov, Ferber, 2014. 8 p. (In Russian).

Percov N.V. O roli korpusov v lingvisticheskih issledovaniyah [About the role of corpuses in linguistic research] Trudy mezhdunarodnoj konferencii "Korpusnaya lingvistika 2006". St. Peterburg: Izd-vo St. Peterb. unta, 2006. 319 p. (In Russian).

Sadovnichij V.A. Bol'shie dannye v sovremennom mire [Big data in the modern world]. Doklad. Moscow, 2017. 4 p. (In Russian).

Yu Yang. Yeesight, Big Data Ecosystem-GTCOM, 2014. 14 p.

Zaharov V.P., Bogdanova S.Yu. Korpusnaya lingvistika [Corpus linguistics]. Irkutsk: IGLU, 2011 (In Russian).

Zubov A.V., Zubova I.I. Informacionnye tekhnologii v lingvistike: Ucheb. Posobie [Information technology in linguistics]. Moscow: Izdatel'skij centr "Akademiya". 9 p. (In Russian).