Программные системы и вычислительные методы
Правильная ссылка на статью:
Скачёва Н.В. Анализ идиом в нейронном машинном переводе: набор данных // Программные системы и вычислительные методы. 2024. № 3. DOI: 10.7256/2454-0714.2024.3.71518 EDN: JLJDSL URL: https://nbpublish.com/library_read_article.php?id=71518

Анализ идиом в нейронном машинном переводе: набор данных

Скачёва Нина Васильевна

ORCID: 0000-0003-2628-6767

старший преподаватель; кафедра лингвистики, теории и практики перевода; Сибирский Государственный Университет Науки и Технологий им. академика М.Ф. Решетнева

660037, Россия, Красноярский край, г. Красноярск, Красноярский рабочий, 31

Sollo_sk@mail.ru

Статья из рубрики "Математическое моделирование и вычислительный эксперимент"

DOI:

10.7256/2454-0714.2024.3.71518

EDN:

JLJDSL

Дата направления статьи в редакцию:

19-08-2024

Аннотация: В разных кругах общественности ни одно десятилетние идут споры о том, может ли «машина заменить человека». Это касается и области перевода. И пока, одни рассуждают, другие «воплощают мечту в реальность». Поэтому сейчас всё больше исследований направлены на усовершенствование систем машинного перевода (далее МП). Чтобы понять преимущества и недостатки систем МП, необходимо, в первую очередь разобраться в их алгоритмах. На данный момент основной открытой проблемой нейронного машинного перевода (НМП) является перевод идиоматических выражений. Значение таких выражений не складывается из значений составляющих их слов, и модели НМП склонны переводить их буквально (т. е. дословно), что приводит к запутанным и бессмысленным переводам. Исследования идиом в НМП ограничены и затруднены из-за отсутствия автоматических методов. Поэтому несмотря на то, что современные системы НМП генерируют все более качественные переводы, перевод идиом остается одной из нерешенных задач в этой области. Это происходит из-за того, что идиомы, как категория многословных выражений, представляют собой интересное

языковое явление, когда общее значение выражения невозможно составить из значений его частей. Первая важная проблема – отсутствие специальных наборов данных для обучения и оценки перевода идиом. В данной работе мы решаем эту проблему, создавая первый крупномасштабный набор данных для перевода идиом. Данный набор данных автоматически извлекается из используемого корпуса переводов с немецкого языка и включает в себя целевой набор, в котором все предложения содержат идиомы, и обычный обучающий корпус, в котором предложения, содержащие идиомы, помечаются. Мы выпустили этот набор данных и используем его для проведения предварительных экспериментов по НМП в качестве первого шага к улучшению перевода идиом.

Ключевые слова:

многословные выражения, идиомы, двуязычные корпусы, машинный перевод, нейронный машинный перевод, немецкий язык, русский язык, лингвистические пары, системы, набор данных

Введение

Нейронный машинный перевод (НМП) в последние годы позволил существенно повысить качество перевода по сравнению с традиционным переводом, основанным на статистическом переводе (SMT) и на правилах и фразах (PBMT). Для понимания вопроса, рассмотрим как работают SMT и PBMT.

Область статистического машинного перевода стала развиваться благодаря появлению большего количества параллельных текстов в лингвистике. Такие параллельные тексты можно встретить в различных лингвистических корпусах. Одним их самых первых и крупных является корпус Europarl. Это сборник протоколов европейского парламента, сформированный с 1996 года и состоящий в то время из 11 языков европейского союза. Этот корпус использовался при создании 110 систем машинного перевода.[1]

Для русского языка важным стало развитие Национального корпуса русского языка (далее НКРЯ), а именно один из его корпусов – параллельный. Который включает в себя такие языки как: английский, армянский, белорусский, болгарский, испанский, итальянский, китайский, немецкий. В корпусе НКРЯ английского языка на данный момент насчитывается 1322 текста и 45 235 028 слов.[2] Сейчас он находится на стадии развития. Большинство систем в значительной степени независимы от языка, и создание системы SMT для новой пары языка в основном зависит от наличия параллельных текстов. Поэтому так важно создавать параллельные тексты в паре с русским языком.

Изначально перевод в SMT строился на словесных моделях IBM.[3] Современные модели уже основаны на фразах. [4] При переводе предложение, фраза исходного языка в любой последовательности слов преобразуются во фразы на целевом языке. Ядром этой модели является вероятностная фраза, полученная из параллельного корпуса. Декодирование представляет собой лучевой поиск по всем возможным сегментациям ввода фразы, любой перевод для каждой фразы и любого изменения её порядка. Учёные пишут о том, что, используя формулу Байеса, можно представить проблему перевода предложения в виде следующего уравнения:

$$\arg\max P(\phi e|\phi r) = \arg\max (P(\phi e) \cdot P(\phi r|\phi e))$$

Uϕe Uϕe

Где □e — это фраза перевода, а □r – фраза оригинала. Поэтому модель языка P(□ является корпусом текста на языке □e, а модель перевода P(□r |□e) – это параллельнь корпус текста на языках ϕr |ϕe.

В таких системах МП в качестве языковой модели используют n-граммы. [5] У Google есть своя коллекция n-грамм, которая является на данный момент самой крупной коллекцией в мире. [6] Самой большой коллекцией n-грамм на русском языке, конечно же является НКРЯ.

Согласно моделям n-грамм, предсказывающее последующее слово после n1.., выявляют возможную последовательность из определенного количества слов. Таким образом есть бигаммы, состоящие из двух слов, триграммы и так далее. Когда количество увеличивается, то подсчитывается вероятность следующего слова или последующих слов. При больших переменных это может выглядеть так:

P(n1,n2,n3,n4) = P(n4|n3,n2,n1)*P(n3|n2,n1)*P(n2|n1)*P(n1)

Рассчитаем вероятность предложения: Ich sehe ein Auto auf der Strasse

P(Ich)

P(sehe|Ich)

P(ein Auto|Ich sehe)

P(Auto|Ich sehe ein)

P(auf|Ich sehe ein Auto)

P(der| Ich sehe ein Auto auf

P(Strasse|Ich sehe ein Auto auf der)

Так как мы чаще встречаем auf der Strasse, чем просто der Strasse, то согласно марковской модели[7] наше предложение будет выглядеть так:

P(Strasse|auf der)

То есть марковская модель n-ого порядка будет выглядеть как:

P(A1 | A1, A2, …, A i-1) ≈P(A1 | Ai-n, Ai-n+1, …, A i-1)

Таким образом, разбитие на n-граммы при переводе текста позволяет найти наиболее удачный перевод текста.

Проблема таких моделей состоит в том, что система не всегда определяет связи между словами, особенно если такие слова стоят далеко друг от друга.

То есть, данные системы МП получили свое распространение благодаря небольшого вложения человеческих ресурсов при условии существования двух параллельных языковых корпусов. Такие системы обучаемы и чем больше текстов находится в параллельных корпусам, тем более адекватнее будет перевод нового текста данной языковой группы. Проблемой для такой системы является перевод идиом, так как они не определяют связи между словами и не чувствительны к контексту.

RBMT – это парадигма МП, в которой лингвистические знание кодируется экспертом в форму правил, которые транслируются из языка оригинала в язык перевода[8]. Данный подход дает полный контроль над производительностью системы, но стоимость формализации необходимых лингвистических знаний значительно выше, чем обучение корпусной системы, в которой процессы происходят автоматически. Однако у данного МП есть свои возможности, даже в условиях невысоких ресурсов.

В RBMT лингвист формализует лингвистические знания в грамматические правила. Система использует такие знания для анализа предложений на исходном языке и языке перевода. Большой плюс такой системы в том, что она не требует каких-либо учебных корпусов, но процесс кодирования лингвистических знаний требует большое количество экспертного времени. Самый крупный проект перевода, построенный на системе RBMT – это Systran.[9]

RBMT – это противоположность систем МП, которые учатся на основе корпусов, представленных нами здесь ранее. Поэтому такая система МП подходит для тех языковых пар, в которых существует мало параллельных корпусов и, как следствие, может охватывать больше языковых пар.

Основной подход системы RBMT основан на связи структуры входного предложения исходного языка со структурой выходного предложения переводимого языка, сохраняя уникальный смысл предложений. Но RBMT также мало восприимчив к переводу идиом.

В чем сложность их перевода? Сложность перевода идиоматических фраз отчасти объясняется сложностью идентификации фразы как идиоматической и создания ее правильного перевода, а отчасти - тем, что идиомы редко встречаются в стандартных наборах данных, используемых для обучения систем нейронного машинного перевода (НМП).

Чтобы проиллюстрировать проблему перевода идиом, мы также приводим результаты работы двух систем НМП для этого предложения в Google и DeepL (см. Таблица 1). Эта проблема особенно ярко проявляется, когда исходная идиома сильно отличается от ее эквивалента в языке перевода, как в данном случае.

Идиома Идиоматический перевод	Es liegt der Hase im Pfeffer! Вот где собака зарыта!
DeepL	Кролик в норе!
Google	Это большое дело!

Таблица 1. Перевод идиомы в системах НМП

Хотя существует ряд моноязычных наборов данных для выявления идиоматических выражений, работы по созданию параллельного корпуса, аннотированного идиомами, необходимого для более систематического изучения этой проблемы, ограничены. Например, американские учёные выбрали небольшое подмножество из 17 английских идиом, собрали 10 примеров предложений для каждой идиомы из интернета и вручную перевели их на бразильско-португальский язык, чтобы использовать в задаче перевода[10].

Создание набора данных для перевода идиом вручную - дорогостоящее и трудоемкое занятие. В этой статье мы автоматически создаем новый двуязычный набор данных для перевода идиом, извлеченных из существующего параллельного корпуса немецко-

русских текстов общего назначения.

Первая часть нашего набора данных состоит из 1500 параллельных предложений, немецкая часть которых содержит идиому. Кроме того, мы предоставляем соответствующие наборы обучающих данных для немецко-русского и русско-немецкого перевода, где отмечены исходные предложения, включающие фразу-идиому. Мы считаем, что наличие большого набора данных для обучения и оценки — это первый шаг к улучшению перевода идиом.

Сбор данных

В данной работе мы сосредоточились на немецко-русском переводе идиом. Для автоматического определения фразеологизмов в параллельном корпусе требуется набор данных, аннотированный вручную лингвистами. Мы используем словарь, содержащий идиоматические и разговорные фразы и созданный вручную, в качестве эталона для извлечения пар идиоматических фраз. При этом обнаружено, что стандартные параллельные корпусы, доступные для обучения, содержат несколько таких пар предложений. Поэтому мы автоматически выбираем пары предложений из обучающих корпусов, в которых исходное предложение содержит фразу-идиому, для создания нового тестового набора.

Обратите внимание, что мы фокусируемся только на идиомах на стороне источника и имеем два отдельных списка идиом для немецкого и русского языков, поэтому мы независимо создаем два тестовых набора (для перевода немецких идиом и перевода русских идиом) с различными парами предложений, выбранными из параллельных корпусов.

Например, в немецком языке подлежащее может находиться между глаголом и предложной фразой, составляющей идиому. Немецкий язык также допускает несколько вариантов перестановки фраз. Чтобы обобщить процесс выявления вхождений идиом, мы видоизменяем фразы и рассматриваем различные перестановки слов в фразе как приемлемое соответствие. Мы также допускаем, что между словами идиоматической фразы может находиться фиксированное количество слов.

Следуя этому набору правил, мы извлекаем пары предложений, содержащие идиоматические фразы, и создаем набор пар предложений для каждой уникальной идиоматической фразы. На следующем этапе мы делаем выборку без замены из этих наборов и выбираем отдельные пары предложений для создания тестового набора.

Для создания новых обучающих данных мы используем оставшиеся пары предложений из каждого набора идиом, а также пары предложений из исходных параллельных корпусов, в которых не было ни одной фразы-идиомы. В этом процессе мы следим за тем, чтобы для каждого идиоматического выражения была хотя бы одна форма как в обучающих, так и в тестовых данных, и чтобы ни одно предложение не было включено как в обучающие, так и в тестовые данные.

При этом для некоторых идиом дословный перевод на язык перевода близок к реальному значению. Пары предложений, в которых идиоматическое выражение использовалось в качестве буквальной фразы будут идентифицированы как идиоматические предложения.

Переводческие эксперименты

Хотя основное внимание в этой работе уделяется созданию наборов данных для обучения и оценки перевода идиом, мы также проводим ряд предварительных

экспериментов НМП с использованием нашего набора данных, чтобы оценить проблему перевода идиом на больших массивах данных.

В первом эксперименте мы не используем никаких меток в данных для обучения модели перевода. Во втором эксперименте мы используем метки в обучающих данных в качестве дополнительной характеристики, чтобы исследовать наличие идиоматической фразы в предложении во время обучения.

Мы проводим эксперимент с немецким и русским языками, предоставляя модели дополнительные входные признаки. Дополнительные признаки указывают, содержит ли исходное предложение идиому, и реализованы в виде специальной дополнительной лексемы , которая добавляется к каждому исходному предложению, содержащему идиому. Это простой подход, который можно применить к любой модели преобразования последовательности в последовательность.

Большинство систем НМП имеют модель преобразования последовательности в последовательность, в которой кодировщик строит представление исходного предложения, а декодировщик, используя предыдущие скрытые элементы LSTM и механизм внимания, генерирует целевой перевод. Мы используем 4-слойную модель кодирования-декодирования на основе внимания, как описывают в работе Тханг Луонг, Хиеу Пхам, Христопхер Д. Маннинг.[11]

Во всех экспериментах словарный запас НМП ограничен наиболее распространенными 30 тыс. слов в обоих языках, и мы предварительно обрабатываем данные исходного и целевого языков, используя 30 тыс. операций слияния.

Мы также используем систему перевода на основе фраз, подобную Moses[12], в качестве базового уровня, чтобы исследовать показатели PBMT при переводе идиом.

Оценка перевода идиом

В идеале перевод идиом должен оцениваться вручную, но это очень дорогостоящий процесс. С другой стороны, автоматические метрики могут быть использованы на больших массивах данных без особых затрат и имеют преимущество в воспроизводимости.

Для оценки качества перевода мы используем следующие метрики, уделяя особое внимание точности перевода идиом: BLEU Традиционная оценка BLEU[13] является хорошим показателем для определения общего качества перевода. Однако эта мера учитывает точность всех n-грамм в предложении и сама по себе не фокусируется на качестве перевода идиоматических выражений.

Модифицированная униграммная точность. Чтобы сконцентрироваться на качестве перевода идиоматических выражений, мы также смотрим на локализованную точность. При таком подходе мы переводим идиоматическое выражение в контексте предложения и оцениваем только качество перевода идиоматической фразы.

Чтобы выделить перевод идиомы в предложении, мы смотрим на выравнивание на уровне слов между выражением идиомы в исходном предложении и сгенерированным переводом в целевом предложении. Для выравнивания слов мы используем функцию fast-align.[14] Поскольку идиоматические фразы и соответствующие переводы во многих случаях не являются смежными, мы сравниваем только униграммы двух фраз.

Обратите внимание, что для этой метрики у нас есть две ссылки: Перевод идиомы как самостоятельного выражения и перевод идиомы, созданный человеком, в целевом предложении.

Точность перевода идиом на уровне слов. Мы также используем другую метрику для оценки точности перевода фразы-идиомы на уровне слов. Мы используем выравнивание слов между исходным и целевым предложениями, чтобы определить количество правильно переведенных слов. Для расчета точности мы используем следующее уравнение:

WIAcc=H-1/N

где H - количество правильно переведенных слов, I - количество лишних слов в переводе идиомы, а N - количество слов в эталонном переводе идиомы.

В таблице 5 представлены результаты для задачи перевода с использованием различных метрик.

Модель	BLEU	BLEU	Униграммная точность	Точность на уровне слов
Базовый уровень PBMT	20,2	19,7	57,7	71,6
Базовый уровень НМП	26,9	24,8	53,2	67,8
Лексема НМП у источника	25,2	22,5	64,1	73,2

Таблица 2. Эффективность перевода на тестовом наборе немецких идиом. Точность идиом на уровне слов и Униграммная точность вычисляются только для фразы-идиомы и ее соответствующего перевода в предложении

Эксперимент НМП с использованием специальной входной лексемы, указывающей на наличие идиомы в предложении, по-прежнему лучше, чем PBMT, но немного хуже, чем базовый вариант НМП, по показателю BLEU. Несмотря на такое падение показателя BLEU, изучая униграммную точность перевода и точность перевода идиом на уровне слов, мы видим, что эта модель генерирует более точные переводы идиом.

Эти предварительные эксперименты подтверждают проблему перевода идиом с помощью нейронных моделей и, кроме того, показывают, что при наличии набора маркированных данных мы можем разработать простые модели для решения этой проблемы.

Выводы

Перевод идиом - одна из самых сложных задач машинного перевода. В частности, было показано, что нейронный МП плохо справляется с переводом идиом, несмотря на его общее преимущество перед предыдущими парадигмами МП. В качестве первого шага к лучшему пониманию этой проблемы мы представили параллельный набор данных для обучения и тестирования перевода идиом для немецкого-русского и русского-немецкого языков.

Тестовые наборы включают предложения с хотя бы одной идиомой на стороне источника, а обучающие данные представляют собой смесь идиоматических и неидиоматических предложений с метками, позволяющими отличить их друг от друга. Мы также провели предварительные эксперименты по переводу и предложили различные метрики для

оценки перевода идиом. Мы формируем новые наборы данных, которые могут быть использованы для дальнейшего изучения и улучшения работы НМП при переводе идиом.

Библиография

1. Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation // School of Informatics University of Edinburgh, Scotland. 2005. P. 79-86.
2. Национальный корпус русского языка. URL: https://ruscorpora.ru/search?search=CgkyBwgFEgNIbmcwAQ%3D%3D (дата обращения 04.03.2024)
3. Brown P. F., Pietra S. A. D., Pietra V. J. D., Mercer R. L. The mathematics of statistical machine translation. Computational Linguistics. 1993. 19(2), p. 263-313.
4. Philipp Koehn, Franz J. Och, and Daniel Marcu.. Statistical Phrase-Based Translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003. p. 127-133. URL: https://aclanthology.org/N03-1017.pdf (дата обращения 05.03.2024)
5. Гудков В. Ю., Гудкова Е.Ф. N-граммы в лингвистике // Вестник ЧелГУ. 2011. № 24.
6. Лингвистический корпус данных. URL: https://catalog.ldc.upenn.edu/byyear (дата обращения 05.03.2024)
7. Жданов А. Е., ДоросинскийЛ.Г. Голосовой замок // Ural Radio Engineering Journal. 2017. Vol. 1, No. 1. P. 80-90.
8. Daniel Torregrosa, Nivranshu Pasricha, Bharathi Raja Chakravarthi, Maraim Masoud, Mihael Arcan. Leveraging Rule-Based Machine Translation Knowledge for Under-Resourced Neural Machine Translation Models // Proceedings of MT Summit XVII, Dublin, 2019, volume 2. URL: https://aclanthology.org/W19-6725.pdf (дата обращения: 19.08.2024).
9. Peter T. Systran as a multilingual machine translation system // Overcoming the language barrier, 3-6 May 1977, Vol. 1. URL: https://www.mt-archive.net/70/CEC-1977-Toma.pdf (дата обращения: 19.08.2024).
10. Salton G., Ross R., and Kelleher J. (2014). An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese // In Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra). 2014. P. 36-41.
11. Luong T., Pham H., Manning C. D. (2015). Effective approaches to attention-based neural machine translation // In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal. 2015. P. 1412-1421.
12. Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. Moses: Open source toolkit for statistical machine translation // In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. 2007. P. 177-180.
13. Papineni K., Roukos S., Ward T., and Zhu W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. 2002. P. 311-318.
14. Dyer C., Chahuneau V., and Smith N. A. A simple, fast, and effective reparameterization of ibm model 2 // In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Atlanta, Georgia, June. 2013 p. 644-646.

Результаты процедуры рецензирования статьи

В связи с политикой двойного слепого рецензирования личность рецензента не раскрывается.
Со списком рецензентов издательства можно ознакомиться здесь.

Статья посвящена исследованию проблемы перевода идиом в контексте нейронного машинного перевода (НМП). Авторы предлагают новый подход к созданию набора данных для перевода идиом на немецко-русских и русско-немецких языковых парах. В статье рассматриваются трудности перевода идиоматических выражений, обусловленные особенностями таких фраз, а также предлагается методология их выявления и перевода.

Методология исследования основывается на автоматическом выявлении идиоматических выражений в параллельных корпусах текстов. Для создания нового набора данных использовался словарь идиом, а также разработаны правила для автоматического выявления и обработки фразеологических единиц. Экспериментальная часть исследования включает сравнение различных подходов к переводу идиом с использованием нейронных моделей и методов статистического машинного перевода (SMT и PBMT). Авторы также предлагают новые метрики для оценки качества перевода идиом.

Актуальность исследования очевидна в условиях растущей популярности и повсеместного использования систем машинного перевода. Проблема точного и адекватного перевода идиоматических выражений является значительным вызовом для существующих систем НМП. Учитывая, что идиомы часто несут важные смысловые и культурные оттенки, разработка эффективных методов их перевода имеет высокую практическую значимость.

Научная новизна работы заключается в предложении оригинального подхода к созданию параллельного корпуса, специально адаптированного для задачи перевода идиом. Авторы представили новый метод аннотирования и автоматического выявления идиом, что является значительным вкладом в область машинного перевода. Также важно отметить предложенные авторами новые метрики для оценки качества перевода идиоматических выражений, которые позволяют более точно оценивать успешность перевода таких фраз.

Стиль изложения статьи академически строг и последовательный, что способствует легкому восприятию сложного материала. Структура статьи логично выстроена: после введения и обзора существующих методов перевода идиом, авторы переходят к описанию методологии исследования, а затем представляют результаты экспериментов. Такой подход позволяет читателю последовательно погружаться в тему исследования и понимать ключевые аспекты предложенных методов.

В заключительной части статьи авторы подводят итоги проведенного исследования, акцентируя внимание на необходимости дальнейшего изучения проблемы перевода идиом и улучшения существующих моделей нейронного машинного перевода. Авторы предлагают перспективные направления для дальнейших исследований, что делает статью полезной как для исследователей в области машинного перевода, так и для практиков, работающих с системами НМП.

Статья представляет собой значимый вклад в область машинного перевода, особенно в контексте перевода идиоматических выражений. Исследование демонстрирует высокий уровень проработки проблемы, оригинальность предложенных решений и их практическую значимость. Рекомендуется к публикации в представленном виде.