СТАТЬИ

УДК 004.912

**ПРИМЕНЕНИЕ МЕТОДОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ
В НЕЙРОННОМ МАШИННОМ ПЕРЕВОДЕ****Ананченко И.В., Полин Я.А., Распопа Е.А., Степанов С.И., Хаджиев И.В.***ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики», Санкт-Петербург,**e-mail: anantchenko@yandex.ru, polin.ya@mail.ru, elizaveta.raspopa.@yandex.ru,
micront@mail.ru, ilya.khadzhiev@gmail.com*

В данной статье рассматривается сравнение различных методов тематического моделирования в контексте использования в системах нейронного машинного перевода. Тематическая модель позволяет определять тематическую принадлежность документа, тематическое моделирование используется для решения задач рубрикации, анализа новостных потоков, в машинном переводе, а также других задачах обработки естественного языка. В сфере нейронного машинного перевода семантический анализ позволяет повысить качество выходных переводимых данных за счет решения задач разрешения синонимии и полисемии терминов, определения функционального стиля и других параметров в процессе анализа текста на исходном языке. Также генерация переводимой информации опирается на принципы распределенной лексической семантики – определении значения слова по свойствам его распределения. В статье рассматривается понятие латентно-семантического анализа – теории извлечения и представления контекстуального значения ряда слов на основе большего количества слов, и вероятностного моделирования – набора алгоритмов лексического анализа корпусных данных с целью извлечения тем и установления связей между ними. В статье приводится сравнение методов классического вероятностного латентно-семантического анализа, робастного алгоритма и латентного размещения Дирихле.

Ключевые слова: машинный перевод, нейронный машинный перевод, нейронные сети, семантический анализ, тематическое моделирование, обработка естественного языка

**APPLICATION OF METHODS OF SEMANTIC ANALYSIS
IN NEURAL MACHINE TRANSLATION****Ananchenko I.V., Polin Ya.A., Rospopa E.A., Stepanov S.I., Khadzhiev I.V.***Federal State Autonomous Educational Institution of Higher education «Saint-Petersburg national
research University of information technologies, mechanics and optics», Saint-Petersburg,**e-mail: anantchenko@yandex.ru, polin.ya@mail.ru, elizaveta.raspopa.@yandex.ru,
micront@mail.ru, ilya.khadzhiev@gmail.com*

The article discusses the comparison of different methods of thematic modeling in the context of the use of neural machine translation systems. Thematic model allows you to determine the thematic affiliation of the document, thematic modeling is used to solve the problems of categorization, analysis of news streams, machine translation and other tasks of natural language processing. In the field of neural machine translation semantic analysis can improve the quality of the output of the translated data by solving the problem of solving synonymy and polysemy of terms, the definition of functional style and other parameters in the analysis of the text in the source language. Also, the generation of translated information is based on the principles of distributed lexical semantics – determining the meaning of the word by the properties of its distribution. The article considers the concept of latent semantic analysis – the theory of extracting and presenting the contextual meaning of a number of words on the basis of a larger number of words, and probabilistic modeling – a set of algorithms for lexical analysis of corpus data in order to extract topics and establish links between them. A comparison of the methods of classical probabilistic latent semantic analysis, robust algorithm and latent Dirichlet placement is presented.

Keywords: machine translation, neural machine translation, neural networks, semantic analysis, topic modeling, natural language processing

На сегодняшний день сфера применения инструментов перевода не ограничивается профессиональной и учебной деятельностью, машинный перевод позволяет расширить спектр доступных данных за счет возможности потребления информации на иностранных языках. Первые системы машинного перевода – на основе правил – фактически имитировали пере-

водческую деятельность человека за счет использования структурированных наборов грамматических правил и лексиконов. Развитие методов анализа и обработки естественного языка спровоцировало переход к математическим инструментам моделирования задачи перевода, таким как *статистический машинный перевод* – метод генерации перевода на основе статистических языковых моделей, параметры которых получены в результате анализа параллельных корпусов исходного и пере-

водимого языков. Распространение технологий нейронных вычислений позволило перейти к более производительным системам *нейронного машинного перевода*, также использующим двуязычные корпуса для построения многомерных языковых моделей. Технология нейронного машинного перевода нуждается в огромном объеме корпусных данных для генерации переводимой информации надлежащего качества, проблемы, с которыми сталкивается нейронный перевод, провоцируют повышение спроса на исследования в сфере обработки естественных языков. Системы нейронного перевода используют векторное представление языка в качестве интерлингвы – промежуточного языка, текст на исходном языке «кодируется» в многомерное представление, а затем «декодируется» – в информацию на переводимом языке. Создание языковой модели предполагает выполнение этапов морфологического, синтаксического и семантического анализа. В контексте перевода методы семантического анализа используются для решения ряда задач: анализа контекста, неологизмов и редкой лексики, проблем синонимии, эллипсов (намеренный пропуск слов, не существенных для смысла выражения) и согласования времен.

Цель работы: теоретическое обоснование применения различных методов тематического моделирования в системах нейронного машинного перевода.

Материалом исследования послужили работы, рассматривающие процессы обработки естественного языка и построения систем машинного перевода.

Одним из факторов, влияющих на качество выходных данных систем машинного перевода, является *функциональный стиль* исходного текста – организация синтаксических и морфологических структур, отвечающих конкретной языковой задаче. Современные системы машинного перевода позволяют успешно обрабатывать тексты, написанные в научном и официально-деловом стиле, данные стили являются разновидностями книжного и характеризуются четкой синтаксической структурой и терминологией, отсутствием средств художественной выразительности. В свою очередь проблема художественного перевода сталкивается с вопросами обработки образов, выраженных с помощью различных языковых средств: эпитетов, метафор, аллегорий, идиом и других структур. Теоретически информация, полученная в результате семантического анализа данных на исходном языке, может послужить инструментом определения стилистической окраски тек-

ста и ограничения рабочего словаря, что, как следствие, может способствовать повышению качества литературного перевода.

Лексемы в контексте нейронной языковой модели представлены многомерными векторами, по одному измерению для каждого слова в словаре, значение 1 присваивается измерению, соответствующему заданному слову. Поскольку подобные векторы являются большим объемом данных, необходимо ввести ограничения на словарь, как автоматические, так и лингвистически обоснованные, к примеру разграничения по частям речи с целью уменьшения размерности векторов. Для нахождения связи между лексемами вводится дополнительный слой между входным и скрытыми слоями, в этом слое каждый элемент контекста индивидуально проецируется в массив данных меньшей разрядности. Использование одной матрицы весов для каждого из контекстных слов позволяет создавать непрерывное пространственное представление для каждого слова, независимое от его положения в предложении. Такое представление называют *векторным представлением слова (word embedding)* – погружение слова в линейное векторное пространство. Таким образом, слова, возникающие в схожем контексте, должны иметь схожие векторные представления. Векторные представления позволяют производить *кластеризацию (clustering)* – обобщение слов, а также предоставлять надежные *прогнозы в случае неизвестного контекста (back-off)*.

Векторные представления фактически являются лексемой, закодированной в многомерном пространстве, векторе около 500–1000 чисел с плавающей точкой, и отражают контекст, позволяя предсказывать следующее слово в последовательности. Наблюдение, что возникающие в одном контексте слова являются семантически схожими, является важным в сфере лексической семантики. Концепт *распределенной лексической семантики (distributional lexical semantics)* заключается в определении значения слова по свойствам его распределения, иными словами – по контексту, в котором оно возникает. Схожесть представлений слов, появляющихся в одинаковом контексте, оценивается с помощью метрики – угла между векторами. При проецировании многомерного представления на двумерную плоскость, схожие слова (к примеру, drama, theater, festival) оказываются кластеризованы [1].

Механизмы *тематического моделирования* позволяют решать, в частности, распространенные в машинном переводе проблемы синонимии и полисемии терминов.

Тематическая модель – модель, позволяющая определять тематическую принадлежность документа или фрагмента текста в коллекции. На вход алгоритма создания тематической модели подается корпус текстов, результатом работы модели является числовой вектор, определяющий степень принадлежности каждого из документов корпуса к рассматриваемой теме. Размерность выходного вектора равна числу тем, данная размерность может быть передана вместе с входными параметрами, либо может быть вычислена моделью самостоятельно. Соответственно, *тематическое моделирование* – процесс построения тематической модели. Для того чтобы решить задачу построения тематической модели, формируется и задается корпус D , причем каждый документ корпуса состоит из последовательности слов, определяемых как $W_d = (w_1, w_2, \dots, w_{n_d})$. Слова текста принадлежат словарию, где W – сам словарь, а n_d – длина некоего документа d . Тему задают комбинации слов и частота их употребления, при этом каждый документ может содержать одну или несколько тем. Построить тематическую модель означает найти темы, которые определены в документе. Это значит, что в процессе формирования такой модели определяются: количество тем в корпусе документов; словарь и распределение частоты употребления входящих в словарь слов, соответствующих определенной рассматриваемой в корпусе текста теме; тематика каждого документа, то есть степень принадлежности текста к каждой из рассматриваемых тем.

Тематическое моделирование включает в себя как задачу кластеризации документов по темам, так и задачу кластеризации слов по темам. При решении чаще всего используют мягкую кластеризацию, когда документ принадлежит нескольким темам, аналогично одно и то же слово может принадлежать нескольким темам [2]. Построение тематической модели позволяет выявлять множество латентных (скрытых) тем и используется для решения дополнительных прикладных задач: разбиение документа на тематические фрагменты; определение тематики творчества автора (если каждый из документов корпуса имеет определенный список авторов); тематический поиск (то есть ранжирование документов по степени соответствия определенной теме); создание тематического каталога документов с определенной иерархией (а также определение правил каталогизации документов); поиск и ранжирование документов в зависимости от степени сходства с некоторым документом.

Во время реализации тематических моделей предполагается выполнение следующих условий: гипотеза о «мешке документов» – иначе называемая *bag-of-documents* – порядок документов в коллекции не имеет значение; гипотеза о «мешке слов» – иначе называемая *bag-of-words* – порядок терминов и текстов в документе не важен; различение и распознавание форм слова как одной лексемы; отбрасываются частотные лексемы – часто встречающиеся предлоги, местоимения, вспомогательные конструкции.

Векторная модель (иначе называют – *векторное пространство модели*, *Vector Space Model*, *VSM*) – представление текстовых документов в алгебраической форме в виде векторов идентификаторов, таких как, например, указатель терминов. Модель используется при фильтрации информации, извлечении информации, индексации и подсчете релевантности рейтинга. Согласно данной модели документ является неупорядоченным множеством термов. Термами в задачах информационного поиска называют слова, из которых состоит текст. Каждому терму сопоставляется его вес в документе, который определяется весовой функцией. Весовые функции основываются на следующих различных статистических методах: *булевский вес* – равен 1, если слово встречается в тексте документа, 0 если не встречается; *term frequency (tf)* – частота вхождения слова в текст документа; *term frequency – inverse document frequency (tf-idf)* – частота вхождения слова в текст документа, деленная на частоту вхождения слова в текст корпуса документов; *log* – величина логарифма от числа вхождения слова в определенный текст.

Латентно-семантический анализ (*Latent Semantic analysis*, *LSA*) – метод, который используется для определения контекстного значения ряда слов на основе большего числа слов. Его используют, когда необходимо найти, проиндексировать или классифицировать документы, а также для построения моделей понимания. Также используется и в других областях, где существует необходимость выявить главные факторы из большого массива информации. LSA использует мешок слов для векторного представления текстовых единиц. Корпус документов обычно представляют в виде матрицы слово-документ. В такой матрице слова, которые используются в корпусе текста, это строки, а столбцы соответствуют документам. Для определения каждого элемента матрицы используют определенные схемы [3]. Например, сингулярное разложение матри-

цы, иначе называемое *SVD* (*singular value decomposition*). Такое разложение показывает геометрическую структуру матрицы. Кроме того, на практике используют некоторые другие особенности сингулярного разложения, например, свойство определения ранга матрицы, приближения матрицы определенного ранга. Принцип LSA заключается в том, что если при использовании некой матрицы A использовать матрицу термины-документы (терм-документная матрица), то в таком случае матрица A' , которая содержит первые k линейно-независимые компоненты первой матрицы (A) будет отображать основную структуру зависимостей в базовой матрице (исходной). Для решения задач тематического моделирования параметр k обычно подбирается экспериментально-эмпирическим путем.

Основным недостатком латентно-семантического анализа является существенное многократное снижение скорости работы алгоритма при увеличении размерности исходной матрицы. Кроме того, некоторые эксперты полагают, что результаты такого матричного разложения не имеют лингвистического обоснования, благодаря чему впоследствии затрудняется оценка и интерпретация результатов.

Вероятностное тематическое моделирование – совокупность алгоритмов, которые дают возможность проводить анализ корпусных данных и извлекать данные о темах корпуса текста и связях между найденными темами. Важным свойством вероятностных тематических моделей является генеративность, поэтому данный тип моделей может быть использован при генерации текстов [4]. Во время реализации такого типа моделирования предполагают, что будет выполняться приведенные ниже гипотезы:

- Каждый термин w в документе d связан с некоторой неизвестной темой $t \in T$.

- Корпус текстовых документов рассматривается как множество параметров (d, w, t) , выбранных случайно и независимо из распределения $p(d, w, t)$, которое задано на множестве $D \times W \times T$. При этом текст $d \in D$ и слово $w \in W$ являются наблюдаемыми переменными, тема $t \in T$ – скрытой.

- Появление слова в документе, относящегося к теме t , описывается общим для всего корпуса распределением $p(w|t)$ и не зависит от документа. То есть, выражая математически – $p(w|d, t) = p(w|t)$. Данная гипотеза называется гипотезой условной независимости.

- Каждый документ и каждое слово w связаны с небольшим количеством тем, поэтому большая часть вероятностей

$p(t|d)$ и $p(w|t)$ должна обращаться в нуль. Эта гипотеза называется гипотезой разреженности.

В рамках каждого документа определяется распределение $\theta_d = p(t|d)$ слов по темам. Таким образом, вероятность обнаружения темы в этом документе становится $\sum_t \theta_{td} = 1$.

При этом каждую из тем можно представить в виде распределения $\phi_t = p(w|t)$ слов из фиксированного словаря, при том, что каждое слово относится к теме с некоторой вероятностью $\sum_w \phi_{wt} = 1$ [5].

Вероятностный латентно-семантический анализ (Probabilistic Latent Semantic Analysis, PLSA) – вероятностная тематическая модель представления текста на естественном языке. Модель основывается на аспектной модели, которая создает соответствия между словами и латентной (скрытой) темой $t \in T = \{t_1, \dots, t_n\}$. Задача тематического моделирования, использующая PLSA, состоит в выявлении этих скрытых переменных. Отличительной особенностью PLSA перед другими моделями, не основанными на вероятностном подходе, является то, что каждый документ в корпусе относится к теме с некой вероятностью, так же слова относятся к теме с некоторой вероятностью. В отличие от LSA вероятностный латентно-семантический анализ корпусов текстов базируется на статистических положениях, что существенно улучшает результаты работы модели. Недостатком PLSA можно назвать то, что повышение числа текстов в корпусе приводит к увеличению параметров, что может привести к чрезмерному переобучению модели. Таким образом, при реализации PLSA при добавлении новых документов в корпус текста необходима полная перестройка и валидация модели.

Латентное размещение Дирихле (Latent Dirichlet Allocation, LDA) – модель моделирования, которая позволяет выявлять причины сходства различных частей данных с помощью объяснения результатов наблюдений через неявные группы. Если рассматривать коллекцию текстовых документов, то можно утверждать, что каждый из них является смещением небольшого числа тем и появление определенного слова в тексте связано с одной из представленных тем в этом документе. Согласно этой модели тематического моделирования, каждый следующий текстовый документ создается независимо от других по определенной схеме [6]: случайно выбирается распределение по некоторым темам θ_d определенного текстового документа; каждое слово в этом

документе в случайном порядке подбирается тема из распределения θ_d и случайно выбирается слово из распределения слов в выбранной теме ϕ_i . При этом в текущем рассматриваемом наборе документов D каждый из них состоит из конкретного числа слов n_d . Известны слова в документе, которые определяются как w_{dn} , в то время как все остальные параметры скрыты. Для каждого документа d переменная θ_d обозначает распределение тем в текущем документе.

Согласно модели LDA число тем в корпусе документов фиксировано и задается параметром T [7]. При этом параметры θ_d и ϕ_i распределены согласно распределению Дирекле по следующему условию: $\theta \sim \text{Dir}(\alpha)$, $\phi \sim \text{Dir}(\beta)$, где параметры α и β – задаваемые векторы, иначе называемые гиперпараметрами. Недостатком распределения Дирекле считают отсутствие лингвистических обоснований [8].

Чтобы оценить качество тематической модели, можно использовать следующие методы: экспертная оценка – также называемая ручной; использование в виде приложения; перплексия (*perplexity*). Последний термин обозначает меру соответствия модели терминам, наблюдаемым в документе и определяемым через логарифм правдоподобия. Меньшая величина значения говорит о том, что модель лучше предсказывает появление слов в документах корпуса. На формуле ниже: $p(w|d)$ – модель, w – термины (слова), находящиеся в документе d принадлежащие корпусу $D(d \in D)$.

$$\text{Perplexity} = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln(p(w|d))\right) \quad (1).$$

Результаты работы

По критерию перплексии робастный вероятностный латентно-семантический анализ превосходит другие рассмотренные методы и является предпочтительным для использования в системах нейронного машинного перевода в процессе построения языковой модели.

Заключение

Значительное распространение систем нейронного машинного перевода обозначи-

ло ряд новых задач, связанных с обработкой естественного языка. Использование статистических методов языкового моделирования не уменьшает значимость глубокого лингвистического анализа, напротив, возвращение к концепции интерлингвы подразумевает анализ входных на всех языковых уровнях (морфологическом, синтаксическом, семантическом, прагматическом), а также многоуровневый синтез текста на переводимом языке. Методы семантического анализа позволяют решать основные проблемы нейронного машинного перевода: разрешение вопросов синонимии, употребления неологизмов, распознавания эллипсов, интерпретации средств художественной выразительности. В работе рассмотрены основные типы тематического моделирования, на сегодня наиболее эффективным методом является вероятностный латентно-семантический анализ.

Список литературы

1. Koehn P. Neural Machine Translation. Center for Speech and Language Processing Department of Computer Science Johns Hopkins University, 2017. 133 p. [Electronic resource]. URL: <http://mt-class.org/jhu/assets/nmt-book.pdf> (date of access: 15.04.2019).
2. Батура Т.В. Семантический анализ и способы представления смысла текста в компьютерной лингвистике // Программные продукты и системы / Software & Systems. 2016. № 4. С. 45–57.
3. Машечкин И.В., Петровский М.И., Царёв Д.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование: новые вычислительные технологии. М.: научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова, 2013. С. 91–102.
4. Jurafsky D., Martin J., Pearson H. Speech and Language Processing. Prentice-Hall Upper Saddle River. 2019. P. 906–908.
5. Landauer T.K., Foltz P., Laham D. An Introduction to Latent Semantic Analysis. DiscourseProcesses. 1998. Vol. 25. P. 259–284.
6. LDA Implemetations [Electronic resource]. URL: <http://www.cs.princeton.edu/~blei/topicmodeling.html> (date of access: 15.04.2019).
7. Гомзин А.Г. Определение тематической направленности текстового содержимого микроблогов // Выходная квалификационная работа: 05.13.00. М.: ВМК МГУ, 2013.
8. Коршунов А.В., Гомзин А.Г. Тематическое моделирование текстов на естественном языке // Труды института системного программирования РАН. 2012. № 23. С. 215–244.