УДК 81’322.4
DOI: 10.18384/2310-712X-2017-5-77-84

МАШИННЫЙ ПЕРЕВОД КАК СРЕДСТВО СТАНДАРТИЗАЦИИ ТЕРМИНОЛОГИИ

Потемкин С.Б.
Московский государственный университет имени М.В. Ломоносова
119991, г. Москва, ГСП-1, Ленинские горы, д. 1, стр. 51, Российская Федерация

Аннотация. Целью статьи явилось описание подхода к стандартизации научно-технической терминологии. Данный подход заключается в прямом и обратном переводе русского термина на иностранный – английский – язык и последующем обратном переводе на русский. В случае совпадения оригинала и результата двойного перевода термина последний предполагается согласованным с международно-принятой терминологией. Расстояние Левенштейна использовалось для определения совпадения / рассогласования между исходным термином и термином, полученным в результате прямого и обратного переводов. Показана возможность выбора русских терминов среди синонимичных или конверсивных словосочетаний. Для оценки семантической близости терминов в предметной области «психология» применён дистрибутивный метод с использованием меры PMI (поточечная взаимная информация). Результаты исследования показывают перспективность использования системы машинного перевода для стандартизации научно-технической терминологии, в частности, в области психологии. Стандартизация в данном случае понимается как согласованность с международной (англоязычной) терминологией.

Ключевые слова: терминология, эквивалент терминов, структура составных терминов, машинный прямой и обратный переводы.

TERMINOLOGY DATABASE AND MACHINE TRANSLATION

S. Potemkin
Lomonosov Moscow State University
1-51, Leninskie Gory, GSP-1, Moscow, 119991, Russian Federation

Abstract. The purpose of this article is to describe the approach to standardization of scientific and technical terminology. This approach consists in the direct and reverse translation of the Russian term into a foreign, specifically, English, language and the subsequent reverse translation into Russian. In case of coincidence of the original and the result of a double translation of the term, the latter is supposed to be consistent with the internationally accepted terminology. The Levenshtein distance was used to determine the coincidence / misalignment between the original term and the term resulting from direct and reverse translation. The possibility of choosing the best Russian term among the synonymous or converse phrases is shown. The distribution method using PMI (Pointed mutual information) metrics is applied for assessing

the semantic proximity of terms. The results of the research show the prospects of using the machine translation system for standardization of scientific and technical terminology, in particular, in the field of psychology. Standardization in this case is understood as consistency with the international, particularly English terminology.

Key words: terminology, term equivalent, structure of composite terms, morphological analysis, machine translation, direct and reverse translation.

Введение

Увеличивающийся объём генерируемой научно-технической информации в мире, который удваивался ежегодно уже в 90-е гг. XX в., требует создания и стандартизации словарей, тезаурусов, энциклопедических справочников, используемых в науке и технике, чтобы позволить читателям правильно понять суть передаваемого сообщения. Это наблюдение в полной мере относится к терминологии, используемой разными авторами в одних и тех же или в смежных областях науки и техники, а также в учебной и научно-популярной литературе. Возможно, ещё более важным вопросом для развития научных исследований является необходимость обмена научной информацией, полученной в разных странах на разных языках. Такой обмен невозможен без адекватного перевода научных статей, в больших объёмах содержащих терминологию.

Проблемы терминологии также связаны с развитием систем машинного перевода. Известно, что автоматический перевод научно-технических текстов при их достаточно бедной синтаксической структуре обусловлен, главным образом, правильным переводом номинативных конструкций, прежде всего терминологических словосочетаний и фраз. На данный момент «уже не вызывает сомнений, что для правильного, научно обоснованного решения терминологических проблем следует изучить терминологию с признанием её природы и логического существования в системе общенародного языка» [4, с. 129]. В рамках вышесказанного упомянутые проблемы терминологии должны изучаться лингвистами и специалистами в данной научно-технической области.

Понятие научно-технического термина

Термин – это слово, фраза, акроним или другая лексическая единица, обозначающая соответствующее экстралингвистическое понятие в реальном мире. По словам М. Глушко, «термин – это слово или фраза для выражения понятий и обозначения объектов, имеющих благодаря своим строгим и точным определениям чёткие смысловые границы и, следовательно, однозначный в рамках соответствующей системы классификации» [5, с. 33]. Как правило, термин должен быть однозначным, т. е. соответствовать только одному конкретному объекту. Уникальность термина, в отличие от слова общей лексики, зависит не только от окружающего контекста. В частности, термин может использоваться изолированно, его принадлежность к определённой терминологии обосновывает уникальность термина в рамках этой терминологии. В другом контексте термин может иметь отличное значение. В научно-технических текстах общеупотребительная лексика часто приоб-

ретает значения, не зафиксированные в обычных словарях. Границу между собственно терминами определённого предметного поля и общеупотребительными словами часто практически невозможно провести [6, с. 168]. Как показывает практика, даже в одной и той же предметной области или в одном и том же информационном поле один и тот же термин может быть поразному определён разными авторами. В качестве примера рассматривается обширная научная область «Психология», выбранная в качестве объекта данного исследования.

Определения, данные разными авторами для термина «фотопсия»:

А) ***фотопсия*** *(от греческого φωτός – свет + ψις – видение) – субъективные световые явления (ощущения), не имеющие характера определённых фигур или объектов. Обычно это мерцающие пятна, искры, светлые зигзаги и т.д. Фотопсия вызвана действием механической или токсической стимуляции зрительного анализатора* [7, с. 780].

Б) ***фотопсия*** *(греческий φωτός – свет + ψις – видение). Появление движущихся фигур, точек, пятен и т.д., обычно светящихся, в поле зрения. Фотопсия наблюдается при заболеваниях сетчатки, а также как элементарные зрительные галлюцинации – психопатические явления* [2, с. 354].

Первое определение не содержит понятия «психопатическое явление», которое играет важную роль в области психологии. Таких примеров множество. Даже один автор часто даёт несколько толкований термина, что отражает различное использование последнего в различных областях науки, а также недостаточную стандартизованность терминологии. Читатель затрудняется в понимании, какой тип объекта обозначен термином в контексте.

Ещё большие трудности возникают также в процессе перевода термина. Например, термин «слияние» имеет, по меньшей мере, 2 английских эквивалента:

СЛИЯНИЕ (симбиотическое) (merging)

СЛИЯНИЕ (синтез) (fusion) [1, с. 211].

Проблема выявления терминов в тексте

Использование сложных терминов – терминологических выражений позволяет частично снять многозначность. В этом направлении была предпринята попытка рассмотреть синтаксическую и логико-семантическую структуру терминов. Очень важный класс составных терминов (MWT, Multi Word Term) – это тот, где значение / семантика термина не выводится из значений составляющих слов. В словосочетании *слепая оценка* отдельные составляющие слова не имеют прямого отношения к фактическому значению синтагмы, которое заключается в опросе экспертов с целью получить суждение о предмете опроса. Напротив, синтагма *слепой человек* не имеет другого толкования, кроме буквального, полученного из композиции составляющих её слов. Подобные переносные значения очень часто встречаются в естественном языке. Обработка MWT подобного типа имеет чрезвычайно высокое значение для адекватной обработки естественного языка [12].

Логическая и семантическая структура термина

Изучение логической и семантической структуры термина выполнялось дистрибутивными методами с использованием метрики поточечной взаимной информации – Pointwise Mutual Information (PMI), определяемой как

PMI (wa, wb) = ln (p(wa, wb) / p(wa) p(wb)),

где wa, wb – термины; p (wa, wb) – вероятность совместного появления wa и wb в документе; p(wa), p(wb) – вероятности появления wa и wb соответственно.

Дистрибутивная гипотеза, хорошо известная в лингвистике, состоит в том, что если слова (словосочетания) встречаются в одних и тех же контекстах, они, как правило, имеют схожие значения. Эта гипотеза является основанием для измерения семантической близости слов. Слово может быть представлено вектором, компоненты которого получены путём подсчёта числа вхождений этого слова в различных документах. Подобие строк-векторов в матрице слова-документы указывает на сходство значений слов.

Мы использовали распознанные и отредактированные тексты журнала «Вопросы психологии» за 1980–2010 гг. [3] в качестве исходного корпуса. Данный ресурс содержит более 13 млрд. словоупотреблений и около 282000 лексем. Статистическое диахроническое исследование этого корпуса опубликовано в 2015 г. [10, с. 95–103]. На его основе эксперты-аналитики выбрали 100 двухсловных терминов в качестве примера терминологии в этой области. Мы выполнили подсчёт встречаемости каждого из этих терминов и PMI для каждой пары терминов [10]. Пары с высоким показателем коллокации имеют высокий PMI, то есть вероятность их совместной встречаемости лишь ненамного ниже вероятности появления каждого термина по отдельности. И наоборот, пара терминов, частота появления каждого из которых намного выше частоты их совместной встречаемости, имеет низкое значение PMI. Самые высокие значения PMI соответствуют наиболее частотной совместной встречаемости пар терминов в статьях журнала. Среди 100 выбранных двухсловных терминов высокие PMI имеют пары:

клиническая психология <> консультативная психология; PMI = 3,6016

ценности жизни <> экзистенциальный анализ; PMI = 4,7447

гендерный анализ <> женская психология; PMI = 4.7729.

Показатель PMI можно использовать для: ранжирования веб-страниц, извлечения редких терминов из текстов ЕЯ, анализа тональности (эмоций) и т. д.

Перевод терминов

На следующем этапе мы провели исследование перевода терминов с помощью системы машинного перевода. Методика этого исследования заключалась в следующем: анализируемый термин подвергался машинному переводу с помощью онлайн-переводчика, в нашем случае Google Translator. Затем выполнялся обратный перевод, и рассчитывалось расстояние Левенштейна [11] между исходным русским текстом и текстом, полученным в результате прямого и обратного переводов [9]. Если это расстояние оказывалось рав-

ным нулю, русский термин признавался согласованным с его иноязычным эквивалентом и, следовательно, соответствующим международным терминологическим стандартам. В противном случае делалось заключение, что система машинного перевода «не знает» русского термина и, соответственно, неправильно выбирает его иноязычный эквивалент. Например,

Initial (Russian) = *вера в справедливый мир*

-> Russian to English = *just-world hypothesis*

-> Back English to Russian = *вера в справедливый мир*

// Levenshtein distance = 0.

В данном случае психологический термин *вера в справедливый мир* согласован с английским эквивалентом *just-world hypothesis.*

В целом более 80% психологических терминов из нашей выборки имели адекватные английские эквиваленты, т. е. можно заключить, что русская психологическая терминология в основном соответствует англоязычной и будет правильно переводиться системой машинного перевода. В то же время ряд русских терминов переводится неверно.

Пример неадекватного перевода термина:

деятельностный подход -> activity approach -> Подход к работе // Levenshtein distance = 3.

Часто перевод терминов требует определённой правки и / или замены на синонимичный вариант. Так, итеративное повторение процедуры прямого и обратного перевода может привести к нулевому расстоянию Левенштейна. Иногда исходный и итеративно полученный термин являются синонимами.

Пример поиска итеративных эквивалентов:

нарушения восприятия времени -> disorders of perception of time-> расстройства восприятия времени -> time perception disorder -> время расстройство восприятия -> time perception disorder -> расстройство восприятия времени.

Можно заключить, что *нарушения восприятия времени* = sin = *расстройство восприятия времени* и английский эквивалент для обоих терминов это *time perception disorder.*

Аналогично синонимами являются термины, полученные в результате прямого и обратного переводов: *запечатление у животных -> imprinting in animals -> импринтинг у животных.*

В других случаях термин, полученный в результате нескольких итераций, является конверсивом исходного: *абсолютная слуховая чувствительность* = conv. = *абсолютная чувствительность слуха;* либо исходный термин и результат итераций имеют различный порядок слов (пермутация) *амнезия раннего детства* = perm = *раннего детства амнезия.*

Следует отметить, что в случае прямого и обратного перевода на язык, отличный от английского, например болгарский, происходит вхождение в бесконечный цикл без получения исходного или семантически близкого термина, расстояние Левенштейна не принимает нулевого значения.

Логический и семантический анализ с использованием машинного перевода позволяет выбирать кластеры связанных терминов с частичным отношением порядка относительно расстояния Левенштейна.

Заключение

Семантическая близость между составными терминами определялась с применением дистрибутивной методики с мерой PMI (поточечная взаимная информация). Дальнейшее развитие этого направления предполагает векторное представление слов (word to vector) [13].

Эквиваленты терминов на английском языке отыскиваются автоматически в процессе прямого и обратного перевода с использованием он-лайн переводчика [8]. Как оказалось, большинство русских психологических терминов (более 80%) адекватно переводятся на английский язык и обратно, т. е. система машинного перевода не встретит затруднений при переводе психологических текстов. В то же время вместо части русских терминов, вероятно, следует употреблять термины синонимичные, которые при переводе на английский язык будут соответствовать общепринятой международной терминологии.

Результаты исследования показывают перспективность использования системы машинного перевода для стандартизации научно-технической терминологии, в частности, в области психологии. Стандартизация в данном случае понимается как согласованность с международной, англоязычной, терминологией.

ЛИТЕРАТУРА

1. Барнесс Э.М., Бернард Д.Ф. Психоаналитические термины и понятия. М.: Класс, 2000. 304 с.
2. Блейхер В.М., Крук И.В. Толковый словарь психиатрических терминов. Воронеж: МОДЭК, 1995. 640 с.
3. Вопросы психологии [Электронный ресурс]. URL: http://www.voppsy.ru/ (дата обращения: 26.05.2017).
4. Гореликова С.Н. Природа термина и некоторые особенности терминообразования в английском языке // Вестник Оренбургского государственного университета. 2002. № 6. С. 129–136.
5. Глушко М.М. Функциональный стиль общественного языка и методы его исследования. М.: Издательство Московского государственного университета, 1974. 120 с.
6. Марчук Ю.Н. Автоматизация перевода и типология текстов // Вестник Московского государственного областного университета. Серия: Лингвистика. 2016. № 2. С. 164–171.
7. Мещеряков Б.Г., Зинченко В.П. Большой психологический словарь. М.: АСТ, Прайм-Еврознак, 2009. 816 с.
8. Онлайн Гугл-переводчик [Электронный ресурс]. URL: https://translate.google.ru/ (дата обращения: 30.05.2017).
9. Потемкин С.Б. Персональный сайт [Электронный ресурс]. URL: http://www.philol.msu.ru/~serge/Translation/form11.php. (дата обращения: 30.05.2017).
10. Потемкин С.Б, Хасин Л.А., Хасина П.Л., Щедрина Е.В. Анализ тенденций развития психологии на основе выявления динамики частоты использования психологических терминов // Вопросы психологии. 2016. № 6. С. 95–103.
11. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. 1965. С. 845–848.
12. Anoop K. Multiword Expression Recognition [Электронный ресурс]. URL: http://pdfs.semanticscholar.org/3e3f/d0173dcb28aa1a11d5342da527a835235ae4.pdf (дата обращения: 15.05.2017).
13. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013. 12 p.

REFERENCES

1. Barness E.M., Bernard D.F. *Psihoanaliticheskie terminy i ponjatija* [Psychoanalytic terms and concepts]. Moscow, Klass Publ., 2000. 304 p.
2. Bleikher V.M., Kruk I.V. *Tolkovyj slovar' psihiatricheskih terminov* [Dictionary of psychiatric terms]. Voronezh, MODEK, 1995. 640 p.
3. *Voprosy psikhologii* [Questions of psychology]. Available at: http://www.voppsy.ru (accessed: 26.05.2017).
4. Gorelikova S.N. [The nature of the term and some characteristics of term formation in the English language]. In: *Vestnik Orenburgskogo gosudarstvennogo universiteta* [Bulletin of Orenburg State University], 2002, no. 6, pp. 129–136.
5. Glushko M.M. *Funkcional'nyj stil' obshhestvennogo jazyka i metody ego issledovanija* [Functional style of public language and methods of its research]. Moscow, MSU Publ., 1974. 120 p.
6. Marchuk Yu.N. [Automation of the translation and typology of texts]. In: *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika* [Bulletin of Moscow Region State University. Series: Linguistics], 2016, no. 2, pp. 164–171.
7. Meshcheryakov B.G., Zinchenko V.P. *Bol'shoj psihologicheskij slovar'* [Big psychological dictionary]. Moscow, AST Publ., Praim-Evroznak Publ., 2009. 816 p.
8. [Online Google translator]. Available at: https://translate.google.ru (accessed: 30.05.2017).
9. Potemkin S.B. [Personal site]. Available at: http://www.philol.msu.ru/~serge/Translation/form11.php. (accessed: 30.05.2017).
10. Potemkin S.B., Khasin L.A., Khasina P.L., Shchedrina E.V. [Analysis of tendencies of development of psychology by identifying the dynamics of the frequency of use of psychological terminology]. In: *Voprosy psikhologii* [Questions of psychology], 2016, no. 6, pp. 95–103.
11. Levenshtein V.I. [Binary codes with correction of deposition, insertions, and substitutions of characters]. In: *Doklady Akademij Nauk SSSR* [Reports of the USSR Academy of Sciences], 1965, pp. 845–848.
12. Anoop K. [Multiword Expression Recognition]. Available at: http://pdfs.semanticscholar.org/3e3f/d0173dcb28aa1a11d5342da527a835235ae4.pdf (accessed: 15.05.2017).
13. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781, 2013. 12 p.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Потёмкин Сергей Борисович – кандидат технических наук, научный сотрудник Центра новых информационных технологий филологического факультета Московского государственного университета им. Ломоносова;
e-mail: potemkin@philol.msu.ru

INFORMATION ABOUT THE AUTHOR

Sergey Potemkin – PhD in Computer sciences, researcher at the Department of New Information technology, Philological faculty of Lomonosov Moscow State University;
e-mail: potemkin@philol.msu.ru

ПРАВИЛЬНАЯ ССЫЛКА НА СТАТЬЮ

Потемкин С.Б. Машинный перевод как средство стандартизации терминологии // Вестник Московского государственного областного университета. Серия: Лингвистика. 2017. № 5. С. 77–84.
DOI: 10.18384/2310-712X-2017-5-77-84