УДК 811.161.1:811.521:004.652.4

ЛИНГВИСТИЧЕСКИЕ И ТЕХНИЧЕСКИЕ АСПЕКТЫ ПОДГОТОВКИ ТЕРМИНОЛОГИЧЕСКОЙ БАЗЫ МНОГОФУНКЦИОНАЛЬНОЙ СИСТЕМЫ ОБРАБОТКИ ТЕКСТОВ ТЕМАТИЧЕСКОГО ДОМЕНА «КАРАТЕ»

С. С. Маевский

Информационно-вычислительный центр Министерства финансов Республики Беларусь, ул. Кальварийская, 17, 220004, г. Минск, Беларусь, maevskiiss@gmail.com

В настоящей статье приведена характеристика терминологической составляющей текстов тематического домена «Карате» как входных данных для многофункциональной системы обработки текстов, дано теоретическое обоснование целесообразности разработки системы подобного типа, проиллюстрирована гипотетическая архитектура системы, а также предложен расширяемый прототип реляционной модели ее терминологической базы.

Ключевые слова: обработка естественного языка; реляционная база данных; терминология; тематический домен «карате».

LINGUISTIC AND TECHNICAL ASPECTS OF PREPARATION OF TERMINOLOGICAL BASE FOR MULTIFUNCTIONAL TEXT PROCESSING SYSTEM OF THE THEMATIC DOMAIN «KARATE»

S. S. Mayeuski

Information and Computing Center of the Ministry of Finance of the Republic of Belarus, Kalvaryjskaja St., 17, 220004, Minsk, Belarus, maevskiiss@gmail.com

This article describes the terminological component of texts of the thematic domain “Karate” as input data for a multifunctional text processing system, provides a theoretical justification for the feasibility of developing a system of this type, illustrates the hypothetical architecture of the system and proposes an extensible prototype of a relational model of its terminological base.

Keywords: natural language processing; relational database; terminology; thematic domain “Karate”.

Карате – японское боевое искусство, разделенное на множество стилей и направлений. Несмотря на призывы известных мастеров прошлого и современности к созданию «единого карате», понимание особенностей отдельных стилей единоборства и даже сути всего боевого искусства в целом зачастую разнится от федерации к федерации, от школы к школе,

от клуба к клубу, от тренера к тренеру. Одним из следствий указанного обстоятельства является различное понимание одинаково именуемых действий и состояний, имеющих место в тренировочном и соревновательном процессах, что находит отражение в текстовых материалах о единоборстве.

Проведенный нами анализ текстовых материалов о карате позволил выявить следующие основные типы текстов, в совокупности составляющих тематический домен «Карате»:

- книги о единоборстве (фундаментальные труды, обучающие пособия, биографии и автобиографии мастеров);
- популярные журналы о единоборстве;
- статьи, посвященные какому-либо аспекту единоборства;
- новостные материалы (афиши, обзоры соревнований, интервью);
- регламентирующие документы (правила соревнований, уставы федераций, турнирные списки, информационные письма и т. д.);
- научно-исследовательские работы (статьи, диссертации);
- стенограммы комментаторской речи на турнирах (в основном субтитры, сгенерированные системами распознавания речи);
- заметки в блогах и социальных сетях.

Из всего многообразия текстов основу тематического домена «Карате», безусловно, составляют фундаментальные труды о единоборстве, обучающие пособия, биографии и автобиографии мастеров, поэтому тексты данного типа составляют важную часть материала для настоящего исследования. Ввиду распространенности карате по всему миру книги о единоборстве издаются на различных языках, однако единой для них остается японоязычная терминология, записываемая средствами алфавита основного языка текста. Осуществленный нами частотный анализ текстов указанного типа выявил периферийность японоязычной терминологии в сравнении с лексемами национальных языков, однако нет никаких сомнений в том, что значительную часть «смыслового ядра» подобных текстов составляет именно японоязычная терминология, подобно тому как французоязычная терминология заключает в себе многие смыслы классического европейского искусства танца, а европейское музыкальное искусство немыслимо без итальяноязычной терминологии.

Во многих случаях японоязычная терминология может быть прямо либо косвенно передана лексемами национальных языков. Так, например, в случае русскоязычной части домена термину *«кидзами цуки»* чаще всего может соответствовать выражение *«прямой удар передней рукой»*, а абстрактному понятию *«дзаншин»* – выражение *«состояние продолженной готовности»*. Указанное соответствие не всегда очевидно, поскольку японоязычные термины могут иметь множество модификаторов,

а также сочетаться с иными терминами. Так, сравнительно короткая тренировочная команда *«дзенкуцу-дачи – кидзами-мае-гери-кэагэ-дзёдан»* средствами русского языка может быть выражена как *«примите основную атакующую стойку с выпадом вперед и нанесите восходящий подбивающий прямой удар передней ногой в верхнюю часть тела»*, но даже такая трактовка не дает неподготовленному человеку исчерпывающего представления о том, что именно требуется сделать (особенно в части постановки тела в правильное положение). С высокой долей вероятности можно предположить, что выводы, касающиеся «русско-японской» терминологии карате, применимы и к множеству других языковых пар.

Для получения представления о текущей лексикографической ситуации для японоязычной терминологии карате, записываемой средствами русского алфавита, нами были проанализированы словари, предлагаемые следующими интернет-ресурсами: *karate.by, kyokushin-iko.by, pobeditel.by, budokan.by, spb-karate.ru, sibkarate.ru, prokimono.ru, mantach-club.ru, kyokushinprofi.com kyokushinkarate.news.*

Материалы существующих словарей терминов карате зачастую представлены в форме простых объединенных по смысловому принципу пар «японоязычный термин – русскоязычный аналог либо толкование», при этом авторы различных словарей придерживаются различного понимания того, к какой смысловой группе отнести тот или иной термин, а также руководствуются различными принципами при включении в словари составных терминов. Попытка применения терминологических словарей для прочтения текстов о карате со всей очевидностью выявляет еще одну закономерность, состоящую в том, что авторы различных материалов придерживаются различных правил передачи японоязычных терминов средствами русского алфавита, нередки также случаи использования различных правил в рамках одного и того же текста. Так, например, упомянутое выше наименование стойки *«дзенкуцу-дачи»* может быть передано как *«дзэнкуцу-дачи»*, *«зенкуцудачи»* (без разделителя), *«дзэнкуцу дати»*, *«дзенкудзу-дачи»* и множеством иных вариантов, формирование которых может быть не ограничено принципами известных русских транскрипций для японского языка, предложенных Е. Г. Спальвиным, Д. М. Позднеевым и Е. Д. Поливановым.

Как упоминалось в начале настоящей статьи, с точки зрения многих мастеров карате позитивной тенденцией для развития единоборства может стать стремление к единству понимания его аспектов. Значительным подспорьем для достижения такого единства может стать разработка многофункциональной системы анализа текстов тематического домена «Карате», которая, в числе прочего, позволила бы проиллюстрировать многообразие написаний и трактовок терминов за счет использования

структурированной терминологической базы реляционного типа. Функциональными элементами такой системы могут стать модуль определения текстов тематического домена «Карате», экстрактор фактов, модуль аннотирования и реферирования текстов, а также иные компоненты, работа которых почти во всяком случае предполагает опору на значительные объемы данных. Пилотная архитектура системы представлена в [1].

Термины карате могут находиться друг с другом подчас в довольно неожиданных взаимоотношениях, а элементы, составляющие термин, имеют различные оттенки значений в зависимости от того, частью какого именно термина они являются. В случае с японоязычными терминами, записываемыми средствами русского (и какого-либо другого) алфавита, проблема активного использования многообразия написания термина в системе обработке текстов не решается «традиционными» способами наподобие вычисления расстояния Левенштейна между условно эталонным и условно вариативным написанием: во-первых, сложность алгоритма вычисления расстояния Левенштейна составляет O(m*n), что очень дорого в плане затрат вычислительных ресурсов, во вторых, применение данного алгоритма не гарантирует точный результат. Так, при классической реализации алгоритма расстояние между вхождениями *«кокуцу-дачи»* и *«дзенкуцу-дачи»* будет меньше, чем между вхождениями *«кокуцу-дачи»* и «кокудзудати», хотя во втором случае сравниваются варианты написания наименования одной и той же боевой стойки. Все перечисленные аспекты должны быть учтены при разработке системы анализа текстов тематического домена «Карате» и найти отражение в структуре ее терминологической базы, пилотная архитектура которой, снабженная подробным описанием назначений и взаимосвязей таблиц, также представлена в [1]. Данная архитектура позволяет осуществлять построение терминов на основании сведений о позициях составляющих их элементов, устанавливать разнообразные отношения между терминами и элементами терминов, ассоциировать термин с вариантами его написания. Несмотря на определенную избыточность, достоинством архитектуры является ее расширяемость: для включения новых таблиц и полей, хранящих сведений о терминах, формируемых средствами нового языка, необходимо совершить действия, полностью аналогичные тем, что уже совершены для включения в архитектуру соответствующих таблиц и полей.

Решением проблемы исчисления вариантов написания терминов может стать алгоритм, входными данными которого являются исходный термин, список символов и комбинаций символов, разделяющих термин на составные части, и список пар «заменяемое–замена». Шаги алгоритма:

1) сортировка входных пар «заменяемое – замена» по длине заменяемых элементов;
2) разделение термина на составляющие части согласно разделителям;
3) последовательное наложение на составляющие термина масок заменяемых элементов, сохранение найденных элементов и их замен в отдельную структуру данных;
4) присовокупление списка разделителей к полученной структуре данных;
5) подготовка вариантов написания термина с помощью декартового произведения множеств, конкатенация.

Прототип терминологической базы реализован нами средствами СУБД SQLite. Также с помощью возможностей Python/Django создано имплементирующее описанный выше алгоритм приложение для работы с прототипом. Рабочий вариант системы, использующей терминологическую базу, планируется к размещению в сети Интернет на сайте www.karatetools.by.

Библиографические ссылки

1. Многофункциональная система обработки текстов тематического домена «Karate» – Пилотная архитектура [Электронный ресурс]. URL: https://karatetools.by/shihan_pilot (дата обращения: 10.03.2024).