

БАЗОВЫЙ СЛОВАРЬ СИСТЕМЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ ЯПОНСКОГО ЯЗЫКА

С. Ю. Кравченко

Минский государственный лингвистический университет

Минск, Беларусь

E-mail: crux57005@tut.by

Предлагается методика и экспериментальные результаты разработки базового словаря системы автоматического лингвистического анализа текстов на японском языке. Пример их успешного применения – лингвистический процессор известной системы автоматизации инженерии и управления знаниями Goldfire Innovator.

Ключевые слова: японский язык, лингвистическая база знаний, базовый словарь.

Одним из основных компонентов лингвистической базы знаний (ЛБЗ) практически любой системы автоматической обработки текстов на естественном языке («базового лингвистического процессора» (БЛП) [1]) является базовый словарь (БС). В БЛП БС используется в составе лингвистического обеспечения алгоритмов автоматической обработки текста – формализации содержания анализируемых текстов, а также снятия лингвистической многозначности объектов и явлений естественного языка (ЕЯ). Данный ресурс включает описание максимально возможного количества единиц анализа (например, символов, иероглифов, слогов, морфем, словоформ или словосочетаний) и их лингвистических признаков, формализованных в соответствии с предварительно выбранной моделью языка [2]. Так, для обеспечения функциональности лексико-грамматического анализа данный ресурс должен содержать максимально возможное количество соответствующих единиц анализа, именуемых «минимальными лексическими единицами» (МЛЕ) с указанием их лексико-грамматических значений – части речи, числа, рода, лица, падежа, времени и залога и т. д. В данной работе под МЛЕ понимаются корневые и аффиксальные морфемы (для записей с использованием иероглифов и слоговых азбук), слова, разделенные символом пробела (например, для записей на английском языке), некоторые экстралингвистические элементы текста – специальные обозначения, части формул, эмодзи и т. д. Для задач синтаксического или семантико-синтаксического анализа текстов БС может быть дополнен информацией о синтаксических и семантико-синтаксических признаках МЛЕ.

Примеры БС для систем анализа текстов на японском языке представлены в [3–6]. Анализ показал их несоответствие целям БЛП, что выражается в громоздкой классификации свойств японского языка, в избыточности и сложности представления словарной информации и, как следствие, в неудобстве работы с ЛБЗ в целом. Поэтому разработка БС японского языка, особенно в целях промышленных систем, весьма актуальной.

Разработку БС для системы автоматического анализа текстов можно разбить на две подзадачи:

- 1) разработка формата БС, отвечающего целям БЛП,
- 2) наполнение БС соответствующим содержанием.

Формат БС должен отражать основные лингвистические характеристики единиц анализа. Что касается «морфемного» подхода к моделированию лексико-

грамматического содержания текста, то достаточно детальная работа по формализации лексико-грамматических свойств японских морфем выполнена в [7]. Ее результаты зафиксированы в виде лексико-грамматического классификатора свойств японского языка, который включил 143 лексико-грамматических кода (ЛГК) для описания лексико-грамматических характеристик всевозможных МЛЕ.

В соответствии с классификатором каждой МЛЕ в тексте может соответствовать один или несколько ЛГК, например:

太陽 NN

где 太陽(ТАЙЁ:, солнце) — это МЛЕ, записанная при помощи кандзи, а NN — ЛГК корня существительного.

С учетом нескольких вариантов записи данного слова при помощи кандзи и азбуки хирагана этот же пример можно представить в виде двух словарных статей:

太陽	NN
たいよう	NN

При этом признак эквивалентности двух МЛЕ можно вынести в специальную словарную статью в самом словаре

太陽 NN = たいよう NN

в которой эквивалентные МЛЕ с их ЛГК записываются в строчку и разделяются знаком равенства (« = »).

Стоит отметить, что слово 太陽 в японском языке может являться и именем собственным (женское имя) и иметь соответствующее произношение и варианты записи (например, ひなた — Хината, или ひかる — Хикару). Данную особенность можно представить следующим образом:

太陽	NN	NP
たいよう	NN	
ひなた	NP	
ひかる	NP	

(где NP — ЛГК корня имени собственного) и добавить соответствующие статьи с указанием эквивалентности этих МЛЕ:

太陽	NN	=	たいよう	NN
太陽	NP	=	ひなた	NP
太陽	NP	=	ひかる	NP

Чтобы описать различные признаки (лексические, грамматические, синтаксические или семантические) единиц анализа, словарные статьи можно разделить на специальные группы. Например, фонетическую и смысловую эквивалентность МЛЕ можно записать следующим образом:

ФОНЕТИЧЕСКАЯ{					
太陽	NN	=	たいよう	NN	
太陽	NP	=	たいよう	NP	
太陽	NP	=	ひなた	NP	
太陽	NP	=	ひかる	NP	
}					
СЕМАНТИЧЕСКАЯ{					
太陽	NN	=	お日様	NN	= 日輪 NN
}					

Для обозначения эквивалентной формы записи МЛЕ в последнем примере был использован символ «=». По аналогии с данной формой записи можно предложить специальную форму для записи словоизменительных характеристик слов, образованных путем присоединения служебных аффиксов к корневым морфемам (символ «→»). Для записи последовательностей МЛЕ (например, корневой морфемы и относящихся к ней аффиксов), их опциональности и вариативности, описательные возможности словарных статей могут быть расширены за счет специального формального языка, функционирующего на уровне слов и словосочетаний. Одним из наиболее известных таких языков является язык расширенных регулярных выражений (РРВ) [8]. Так, например, словарная статья, описывающая формирование отрицательной формы глагола в японском языке, записанная при помощи РРВ, будет иметь следующий вид:

<1 VB1S > SV1SU → \$1 さ_SV1SA な_SPNEG い_SPI

Как видно из примера, для изменения формы глагола с утвердительной (話す — ХАНАСУ, разговаривать) на отрицательную (話さない — ХАНАСАНАЙ, не разговаривать), могут быть использованы специальные шаблоны для ЛПК (запись, расположенная слева от символа «→»), а также соответствующие операции по генерации производной формы (запись, расположенная справа от символа «→») — копирование корневой морфемы (話), выделенной треугольными скобками в левой части статьи, при помощи знака \$ с указанием номера треугольной скобки в правой части статьи, а также генерация соответствующих данной форме суффиксальных морфем (さない).

Основным отличием данного способа представления БС от существующих словарей является выделение каждого из лингвистических признаков МЛЕ в отдельную словарную статью, что упрощает представление обширной лингвистической информации, позволяет дополнять и изменять БС, сосредотачиваясь на интересующих эксперта аспектах. Таким образом, работая, например, с автоматическим морфологическим анализом текста, лингвист-эксперт сможет быстро получить доступ к интересующей его информации, изменить или дополнить ее, не отвлекаясь на другие признаки единиц анализа.

Подзадача по наполнению БС содержанием не менее важна, чем разработка формата словарных статей, поскольку именно объем и качество БС обуславливают качество работы всего БЛП. Даже при ограничении тематики БС требования к данному ресурсу остаются высокими, чтобы обеспечить максимально качественный анализ постоянно появляющихся новых текстов [1]. Важно помнить, что лексикон любого живого языка, как и сам язык в целом — постоянно развивающаяся и меняющаяся система, и не существует такого словаря, который мог бы полностью включить все лексические единицы какого-либо активно используемого языка. Тем не менее динамика развития некоторых групп слов позволяет говорить об их устойчивости. В связи

с этим для решения задачи наполнения БС имеет смысл разделить все единицы анализа и их характеристики на две группы:

- открытая группа, т. е. такая группа лингвистических объектов, полное покрытие которой в любом словаре обеспечить практически невозможно, как по причине ее слишком большого объема, так и по причине ее постоянного пополнения неологизмами и заимствованиями;

- закрытая группа, т. е. такая группа лингвистических объектов, которая содержит конечный и неизменный набор элементов.

С точки зрения задачи по наполнению БС, к открытой группе лингвистических объектов японского языка можно, как показывает анализ, отнести словарные статьи следующего типа:

- имена существительные японско-китайского происхождения (МЛЕ с ЛГК NN и NP),

- слова-заимствования и неяпонские имена собственные, записываемые при помощи азбуки катакана (например, カーディオオーバーター – КА:ДИОБА:ТА:, кардио-вертер),

- условные обозначения, сокращения, формулы и эмотиконы (например, H₂O, DVD, :-D, ^_^ и т. д.),

- признаки семантической эквивалентности.

К закрытой группе относятся:

- служебные слова и морфемы, традиционные японские глаголы, прилагательные и наречия (по сути, к данной группе относятся все МЛЕ, которые не являются частью открытой группы),

- признаки фонетической и графической эквивалентности МЛЕ,

- словоизменительные свойства лексических единиц.

Как отмечается в работе [2], при разработке БЛП, первоочередным источником словарной информации для БС является базовый корпус текстов (далее БКТ). Однако в этой же работе делается вывод о том, что БКТ – лишь основа для БС, и данный ресурс не способен предоставить достаточное количество словарной информации – работа по наполнению БС должна продолжаться за рамками задачи по разработке БКТ.

Как выявлено в результате работы по наполнению предложенного БС содержанием, для наибольшей эффективности наполнение БС должно производиться последовательно, в несколько этапов:

1. Разработка БКТ и пополнение БС информацией из корпуса.

2. Наполнение содержанием всех закрытых групп БС усилиями лингвистов-экспертов.

3. Наполнение открытых групп БС при помощи доступных электронных ресурсов (толковых и переводных словарей, списков специальных терминов) с минимальным участием лингвистов-экспертов (валидация добавляемой информации).

4. Наполнение открытых групп БС при помощи специально созданных для этого программ — полностью автоматически или с минимальным участием лингвистов-экспертов (валидация добавляемой информации).

Последний этап требует более подробного описания, так как именно автоматизация получения словарной информации является одним из наиболее эффективных методов ее сбора. Не существует универсального алгоритма, позволяющего получить словарную информацию для открытых групп всех типов — каждая проблема требует своих решений. Например, для получения списка имен существительных японско-китайского происхождения можно предложить полуавтоматический подход, включающий автоматическое выделение слов-кандидатов из корпуса текстов на основа-

нии контекста и последующую валидацию выделенных слов усилиями лингвистов-экспертов. Полностью автоматический алгоритм можно предложить для добавления новых слов-заимствований, записанных при помощи катакана [9].

Что касается условных обозначений, сокращений, формул и эмотиконов, то образование новых слов осуществляется на основе специальных символов по особым правилам. Хранение в БС всевозможных выражений данного типа не представляется возможным (хотя хранение их небольшой части может оказаться полезным), поэтому более эффективным способом описание набора шаблонов на основании специального формального языка, в результате срабатывания которых можно было бы отнести то или иное слово к данной группе. Примером такого ресурса может служить работа, посвященная анализу химических формул и выражений [10].

Задача по выделению признаков семантической эквивалентности является задачей поиска синонимов и является одной из наиболее актуальных проблем современной компьютерной лингвистики. Для решения данной задачи могут использоваться как специальные словари синонимов, онтологические словари, тезаурусы, так и программы автоматического обнаружения синонимичных слов и выражений [11].

Используя формат БС и методики его пополнения, группе из 3 лингвистов-экспертов за два месяца удалось получить результаты, показанные в таблице.

Содержание базового словаря

Компонент БС	Размер (в уникальных МЛЕ)
Лексико-грамматические признаки МЛЕ (МЛЕ и все допустимые ЛГК)	253769
Семантические признаки МЛЕ (существительные времени, процессов, параметров, имена, фамилии, названия; глаголы движения, говорения, переходные и непереходные глаголы и т. д.)	8542
Признаки семантической эквивалентности	12625
Признаки графической и фонетической эквивалентности	8526
Словоизменительные характеристики	3399

Как показали реализованные приложения, размера полученного БС достаточно для обеспечения качественной функциональности БЛП.

Библиографические ссылки

1. Чеусов А. В. Разработка лингвистических процессоров промышленной обработки текстовых документов // Искусственный интеллект. Минск : 2006. № 4. С. 635–639.
2. Воронцов А. В. Лингвистический процессор: его архитектура и особенности функционирования. // Информационные системы и технологии (IST'2006): материалы III Междунар. конф. Минск : Академия управления при Президенте Республики Беларусь, 2006. Ч. 2. С. 27–33.
3. EDR. ELECTRONIC DICTIONARY VERSION 1.5 TECHNICAL GUIDE. Technical report // Japan Electronic Dictionary Research Institute Ltd. 1996. <http://www2.nict.go.jp/r/r312/EDR/index.html>.
4. Курохаси С. Система морфологического анализа японского языка JUMAN. Версия 6.0. // Аспирантура кафедры информатики университета Киото. 2009.
5. Kawata Y., Bartels J. Stylebook for the Japanese Treebank in VERBMOBIL // Verbmobil-Report 240. 2000. ISSN 1434–8845.
6. Asahara, M., Matsumoto Y. IPADIC version 2.7.0 User's Manual // Computational Linguistics Laboratory. Graduate School of Information Science. Nara Institute of Science and Technology. November. 2003.

7. *Кравченко С. Ю.* Лексико-грамматический классификатор японского языка // Вестн. МГЛУ. Сер. 1, Филология. № 3(58). Минск, 2012. С. 108–116.
8. *Cheusov A.* The Word-based Regular Expressions in Computational Linguistics // Proc. of the seventh International Conf. «Pattern recognition and information processing». Minsk, 2003.
9. *Knight K., Graehl J.* Machine transliteration // Computational Linguistics. MIT Press. 1998. № 24. P. 599–612.
10. *Townsend J., Copestake A., Murray-Rust P., Teufel S., Waudby C.* Language Technology for Processing Chemistry Publications // Proceedings of the fourth UK e-Science All Hands Meeting (AHM-2005). Nottingham, UK: 2005.
11. *Wang T., Hirst G.* Exploring patterns in dictionary definitions for synonym extraction // Journal of Natural Language Engineering. 2012. № 18.