EURASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
Innovative Academy Research Support Center
UIF = 8.3 | SJIF = 5.916 www.in-academy.uz

СОСТАВЛЕНИЕ СЛОВАРЯ: ТОКЕНИЗАЦИЯ СЛОВ

[1]Ахмедов Э.Ю
MSc, ассистент, кафедра программный инжиниринг, Ургенчский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хорезмий, " 05.00.00 - Технические науки", +998974567912,
[2]Хаитбоева Д.З
MSc, ассистент, кафедра программный инжиниринг, Ургенчский филиал Ташкентского университета информационных технологий имени Мухаммада ал-Хорезмий, " 05.00.00 - Технические науки ", +998940599050.
https://doi.org/10.5281/zenodo.7493692

ARTICLE INFO

Received: 20 th December 2022
Accepted: 28th December 2022
Online: 30th December 2022

KEY WORDS

Корпус текстов, верстка словаря, словари морфологического анализа, токенизация, словарь – лексикон, синтаксический анализатор – компилятор, токен, терм, слово, n-грамма - токен, символ или терминальный символ.

ABSTRACT

Задачи, стоящие перед современной лингвистикой, выдвинули на первый план использование компьютерной техники для автоматической обработки текстов. Задача использования современных компьютерных технологий в лексикографии остается по-прежнему актуальной и требует детального изучения. Современные вычислительные средства дают возможность автоматизировать лексикографическую работу практически на всех этапах - от выбора цитат до редактирования словаря и его печати.

Автоматизация рутинных процедур, широкое распространение компьютерных программ повышают производительность труда отдельного лексикографа. В результате этого в рамках лексикографии сформировалось новое направление - компьютерная лексикография, включающая в себя создание автоматических словарей, а также разработку программ поддержки лексикографических работ.

К основным направлениям компьютерной лексикографии можно отнести:

1) автоматическое получение из текста с помощью компьютерных средств различных словарей (частотных, терминологических, конкордансов и т.д.);

теоретические и практические аспекты составления компьютерных словарей для систем обработки естественного языка (Natural Language Processing);

2) создание и эксплуатация словарей, являющихся машинными версиями традиционных словарей.

Первое и второе направления занимаются разработкой программ поддержки лексикографических работ.

В настоящее время весьма актуальным является ввод на машинные носители известных словарей и справочников и создание на их базе новых словарей. Перевод в машинную форму ранее опубликованных книжных словарей позволяет «разложить по полочкам» содержимое каждого из них. Только на этой основе и можно осуществить

эффективный контроль полноты и последовательности заполнения полей в статьях словаря, а также эффективно использовать и контролировать информацию в последующей лексикографической работе над новыми версиями данного словаря.

Методы компьютеризации лексикографических работ. Электронные словари

Меняется инструментарий науки, создаются новые словарные технологии, изменяется содержание труда лексикографа. Постепенно традиционные методы заменяются компьютерной обработкой лексикографических данных.

Традиционная технология создания словаря выглядит так:

Формирование словника словаря - Поиск примеров и формирование картотеки примеров - Написание словарных статей - Создание рукописи словаря - Перепечатка рукописи - Редактирование словарных статей - Авторская доработка - Перепечатка рукописи - Корректура - Перепечатка рукописи - Набор, верстка словаря - Корректура - Печать словаря - Словарь.

Компьютерная технология создания словаря включает в себя следующее:

Формирование корпуса текстов - (Создание словника) - Автоматическое формирование корпуса примеров - Написание словарных статей - Ввод словарных статей в базу данных - Редактирование словарных статей в базе данных - Корректура текста в базе данных - Порождение текста словаря и формирование оригинал-макета - Печать словаря - Словарь.

Электронный (автоматический) словарь - это словарь в специальном машинном формате, функционирующий как часть программного обеспечения компьютера. Сегодня широко распространяются электронные версии самых различных словарей. В отличие от традиционных словарей электронный словарь наряду с текстом и графическими изображениями может содержать весь спектр медиа объектов, включая видео и анимационные фрагменты, звук, музыку и прочее.

Все электронные словари можно разделить на два типа:

1) автоматические словари конечного пользователя (о них и пойдет речь в данном разделе);

2) Автоматические словари для программ обработки текста (это информационно-поисковые тезаурусы, частотные словари, рубрикаторы, классификаторы, словари морфологического анализа; словари для машинного перевода), которые включают подробную информацию о морфологических, синтаксических и семантических особенностях функционирования слова.

Важной особенностью электронного словаря является его гипертекстовое устройство. Ссылки, внедренные в слова, фразы или рисунки, позволяют пользователю выбрать текст или рисунок и немедленно вывести на экран связанные с ним сведения и мультимедийные материалы. Взаимоотношения между компонентами словарной статьи не являются линейными. Словарная статья имеет четкую логическую структуру с иерархическими связями между элементами. Каждая информационная категория занимает здесь строго фиксированное место - так называемую «зону».

Электронные словари имеют серьезные преимущества по сравнению со своими бумажными аналогами, что проявляется в быстром росте соответствующего рынка.

Электронный словарь принципиально может обойти ключевое противоречие книжной лексикографии: чем больше информации предлагает словарь, чем больше развит его научный аппарат, тем сложнее им пользоваться. Поэтому классические словари разделяются на две категории. Первая - популярные, относительно удобные, но довольно простые. Вторая - обстоятельные академические издания, не всегда позволяющие быстро получить искомую информацию. Современные электронные словари не только значительно превосходят по объему книжные, но и находят искомое слово или словосочетание за несколько секунд.

Фактически многие словари, которые сформировались в языковой атмосфере середины прошлого века, сильно устарели. Появляются новые отрасли производства, науки, бизнеса, культуры. В обычную разговорную речь приходят новые слова, термины, устойчивые словосочетания. В них не указаны современные значения старых слов, а многие новые слова просто отсутствуют, так как бумажные словари слишком долго готовятся. Электронные словари могут оперативно обновляться.

Очень важно, что электронные словари используют последние достижения лексикографии. Каждое значение в электронном словаре сопровождается синонимами, антонимами, примерами употребления, лингвистической информацией.

Однако разработка электронных словарных баз, так же, как и создание бумажных словарей, является трудоемким делом, а лицензирование готовых словарей обходится очень дорого.

На сегодняшний день получили широкое распространение электронные словари разных издательств: Lingo (ABBYY Software House), Мультиплекс (МедиаЛингва), Polyglossum (ЭТС - "Электронные и традиционные словари"), Контекст (Информатик), PROMT (ПРОМТ) и многие другие. Эти словари в большой степени универсальны, но вместе с тем каждый из них тяготеет к определённой нише.

Два самых известных электронных словаря - Lingo компании Abby и Мультиплекс, разработанный фирмой МедиаЛингва. Специалисты, создающие эти словари, исповедуют разные взгляды на принципы электронной лексикографии. Компания МедиаЛингва придерживается при разработке словарей мультиплекс стратегии, которая заключается в создании цифровых копий известных книжных изданий.

Подход МедиаЛингва имеет и недостатки, так как жесткая привязка к бумажному прототипу не дает возможности исправлять и дополнять электронный словарь, а тем более изменять структуру построения словарной статьи. Традиционные словари довольно серьезно отстают от языковой реальности - обычно это не менее десяти лет. А электронные словари можно пополнять чуть ли не ежедневно [1].

На сегодняшний день в магазинах, специализирующихся на продаже программного обеспечения, можно найти целый ряд словарей и энциклопедий на компакт-дисках. Конечно, словари и энциклопедии широко представлены и в российском Интернете - от крупных проектов до не менее интересных тематических словарей, созданных и пополняемых энтузиастами.

Один из самых популярных порталов онлайн-словарей в российской Сети - ИПС flndex. Словари (http://slovari.yandex.ru/). Портал осуществляет перевод как на, так и с английского, немецкого, французского, итальянского, испанского, украинского языков.

Помимо словарей общей лексики, предлагаются справочники медицинские, юридические, технические и многие другие. Портал очень прост и удобен в использовании.

Портал Mail.ru (http://multilex.maii.ru/) представляет словари на семи языках. Онлайн-справочники включают в себя как общую лексику, так и медицинскую, экономическую, а также словари терминов нефтяной, газовой отраслей, солнечной энергетики и другие отраслевые словари.

Существуют онлайн-словари и на ИПС Rambler - Словари (http://www.rambler.ru/dict/). Выбор языков на портале пока небольшой: английский и немецкий.

Сейчас все более популярными становятся электронные словари на карманных персональных компьютерах (SlovoED/Multiplex, Abbyy Lingvo, Pocket Context, Absolute Word Roadlingua, Diet, VVS Словарь, Pocket Multitran, Pocket Promt и т.д.) [2].

В NLP токенизация - это особый вид сегментирования документов. При сегментации текст разбивается на мелкие куски (сегменты) с более узким информационным содержанием. Сегментация может включать разбиение документа на абзацы, те на предложения, их на фразы и последние - на токены (слова), а также знаки препинания.

Вот эквиваленты основных блоков NLP в компиляторах языка программирования:

- токенизатор - сканер, лексический анализатор;
- словарь - лексикон;
- синтаксический анализатор - компилятор;
- токен, терм, слово или n-грамма - токен, символ или терминальный символ.

Токенизатор разбивает неструктурированные данные, текст на естественном языке, на фрагменты информации, которые можно считать отдельными элементами.

Самый простой способ токенизации предложения - применение внутри строк пробелов в качестве разделителей слов. В языке Python для этого подходит метод split из стандартной библиотеки, доступный для всех экземпляров объекта str, а также для самого встроенного класса str (листинг 1).

Листинг 1. Токенизации предложения

```
>>> stroka = " Сильные порывы ветра срывают с деревьев жёлтые листья 26."
>>> stroka.split()
[‘Сильные’,
'порывы'
'ветра',
‘срывают’,
‘с’,
‘деревьев’,
‘жёлтые’,
‘листья.’
’26.’]
```

Этот встроенный метод языка Python неплохо токенизирует простое предложение. Его единственная «ошибка» заключается в последнем слове, где он включил в токен 26., завершающий предложение, знак препинания. Обычно токены должны быть отделены от соседних знаков препинания и других значимых токенов в предложении. Токен 26. - прекрасное представление числа с плавающей запятой 26.0, но при этом он отличается от числа 26, встречающегося в корпусе в середине предложения, или 26? в конце вопросительного предложения. Хороший токенизатор должен удалить этот дополнительный символ, чтобы создать 26 в качестве класса эквивалентности для 26, 26!, 26? и 26.

Более точный токенизатор создает отдельный токен для любого завершающего предложение знака пунктуации, чтобы сегментатор предложения или детектор границ могли найти конец этого предложения [3].

References:

1. Селегей, В. Электронные словари и компьютерная лексикография / В-. Селегей // Ассоциация лексикографов Lingvo [Электронный ресурс]. Режим дос¬тупа: http://ww'w.lingvoda.ru/translorum/articlcs/selegey_al asp. Дата доступа: 15.09.2009.
2. Герд, А.С. Прикладная лингвистика / А.С. Герд; С.-Петерб. гос. ун-т. - Санкт-Петербург: Изд-во С - Петербургского университета, 2005. 268 с.
3. Обработка естественного языка в действии. - СПб.: Питер, 2020. - 576 с.: ил. -(Серия «Для профессионалов») ISBN 978-5-4461-1371-2
4. Математическая лингвистика и автоматическая обработка текстов: учеб. пособие / Т. В. Батура; Новосиб. гос. ун-т. – Новосибирск: РИЦ НГУ, 2016. – 166 с. ISBN 978-5-4437-0548-4