И. А. Меньшенина (*Витебск, УО ВГМУ*)

ФОРМАЛИЗАЦИЯ РАСПОЗНАВАНИЯ КОМПОНЕНТА «РЕЗУЛЬТАТЫ» В НАУЧНЫХ МЕДИЦИНСКИХ СТАТЬЯХ НА АНГЛИЙСКОМ ЯЗЫКЕ

В работе представлены данные анализа раздела «Результаты» англоязычных научных статей по медицине. Выделяются прототипические компоненты суперструктуры данного раздела с выявлением наиболее употребительных языковых маркеров для каждого из них. Предлагается алгоритм поиска семантических компонентов раздела «Результаты» медицинского научного текста.

К л ю ч е в ы е с л о в а: медицинская научная статья; суперструктура; семантический компонент; языковые маркеры; автореферирование.

In this work English scientific articles in medicine are analyzed. The prototypical components of the semantic structure are discussed and their linguistic markers are identified. An algorithm to identify the semantic components of the “Results” section of a medical scientific text is offered.

K e y w o r d s: scientific medical article; superstructure; semantic component; autosummarization; linguistic markers.

Одной из самых серьезных проблем современного общества является стремительное увеличение объемов текстовой информации, которую должен воспринимать и использовать человек в процессе своей профессиональной деятельности. Для специалистов медицинского профиля этот вопрос особенно актуален, так как от скорости и верности их решений зависит здоровье и жизнь пациентов. Особую роль в решении данной проблемы приобретают системы автоматического реферирования текстов, которые извлекают наиболее важные сведения из одного или нескольких документов и генерируют на их основе лаконичные отчеты, в которых сохраняется смысл оригинала.

Целью настоящей работы является создание алгоритма автоматического определения основного содержания текста научной медицинской статьи, используя когнитивно-дискурсивный подход.

Несмотря на наличие макроструктуры в тексте научной медицинской статьи, которая эксплицитно манифестируется при помощи заголовка, подзаголовков, оглавления, композиционного членения и т.п. [1, с. 136], важность информации в большинстве случаев обусловлена смысловой структурой текста как представителя определенного жанра. Таким образом, одной из задач настоящего исследования является определение закономерностей структурно-семантической организации текста научной медицинской статьи, что позволит конкретизировать наиболее важные смысловые компоненты данного типа текстов для последующего включения их в реферат, а также выделение лексических и грамматических маркеров, при помощи которых смысловые сегменты вербализуются в тексте.

Разделы научной медицинской статьи «Введение» и «Методы» были описаны нами ранее [2; 3]. В данной статье анализируется смысловая структура раздела «Результаты». Материалом для исследования послужили 100 статей из ведущих англоязычных медицинских журналов «The Lancet», «The British Medical Journal (BMJ)», «The New England Journal of Medicine», «The Journal of the American Medical Association» и др. Отобранные статьи были опубликованы в период с 2015 по 2020 годы.

Раздел «Результаты» является, по сути, квинтэссенцией (ядром) всей статьи, поскольку дает ответ на цель или гипотезу, поставленную автором. Автор, как правило, избегает комментариев и интерпретаций, приводя лишь сухие факты, выраженные в цифрах. Данный раздел научной медицинской статьи отличается широким использованием невербальных средств, таких как рисунки, таблицы, диаграммы и графики.

В этом разделе в полной мере реализуются все знаковые особенности научной медицинской статьи: объективность, четкость, безэмоциональность, логичность и т.п. Автору необходимо максимально объективно и полно описать полученные результаты, не обходя вниманием и те, которые спорят с выдвигаемой гипотезой.

При анализе смысловой структуры данного раздела были выделены следующие семантические компоненты:

1) описание материала исследования;
2) основные результаты исследования;

3) дополнительные результаты исследования;
4) нежелательные явления.

Алгоритмический поиск фрагментов текста, соотносимых с семантическими компонентами раздела «Результаты» возможен при определении способов их маркирования в тексте. С помощью методов анализа поверхностной структуры текста, таких как метод сигналов (bonus words, stigma words и т.д.), метод индикаторных фраз, позиционный метод и т.п., были определены маркеры, используемые для вербальной реализации семантических сегментов раздела «Результаты».

Рассмотрим более подробно компонент «описание материала исследования».

Здесь автор представляет исходные демографические и клинические характеристики выборки, а также поэтапно описывает схему проведения исследования.

Описание участников исследования включает пол, возраст, показатели физического развития, клинико-анамнестические признаки для формирования у читателя четкого представления об исследуемой группе, что, в свою очередь, позволяет судить о степени обобщаемости выводов.

Схема исследования обычно включает информацию о количестве скринированных пациентов; о количестве включенных в исследование пациентов (с разделением по сравниваемым группам); о количестве получивших исследуемый медицинский продукт или услугу; о количестве завершивших исследование; о количестве участников, включенных в анализ, а также о причинах, по которым скринированные пациенты не включались в исследование, а выбывшие участники прекращали свое участие в нем [4].

Следовательно, для вербализации данного компонента используются маркеры *men* ‘мужчины’, *women* ‘женщины’, *the median age/time* ‘средний возраст/время’, *at enrolment* ‘на момент регистрации’, *demographic (and clinical) characteristics* ‘демографические (и клинические) характеристики’, *eligibility criteria* ‘критерии отбора’, *mean age* ‘средний возраст’, *exclusion criteria/criterion* ‘критерии исключения’ и т.п., а также лексические индикаторы *We enrolled* ‘мы зарегистрировали’, *we assigned* ‘отобрали’, *we recruited* ‘набрали’, *were assigned to* ‘были отобраны для’, *were enrolled* ‘были зарегистрированы’, *were included* ‘были включены’ и т.п.

Demographic and clinical characteristics *were similar in the two treatment groups across cohort 1 and cohort 2.* ***The mean age*** *was 72·5 years (SD 14·6),* ***women*** *comprised 48 % of the participants.* ‘Демографические и клинические характеристики были сходны в двух группах лечения в когорте 1 и когорте 2. Средний возраст составил 72,5 лет, женщины составили 48 % участников’.

Следующий семантический компонент описывает основной исход исследования и связанные с ним результаты статистической обработки данных. Анализ особенностей данного компонента обнаружил достаточно однородную картину в способах его вербализации: *primary endpoint* ‘первичная конечная точка’, *main results* ‘основные результаты’, *net clinical outcome* ‘чистый клинический результат’ и некоторые другие. Приведем примеры:

***The primary outcome**, an ordinal comparison of the distribution of patients across the mRS categories at 6 months, adjusted for variables included in the minimisation algorithm, was similar in the two groups.* 'Первичный результат – порядковое сравнение распределения пациентов по категориям mRS через 6 месяцев с поправкой на переменные, включенные в алгоритм минимизации, – был сходным в обеих группах'.

Назначение семантического компонента «Дополнительные результаты исследования» заключается в описании дополнительных исходов исследования, которые, как правило, предварительно сформулированы в предыдущем разделе статьи. Данный раздел также часто включает результаты оценки эффекта медицинского вмешательства в подгруппах [4, с. 428].

Выделить данный семантический компонент позволяют такие языковые маркеры, как *secondary endpoints* 'вторичные конечные точки', *secondary outcomes* 'второстепенные ожидаемые результаты', *additional efficacy outcomes* 'вспомогательные конечные показатели'.

*With regard to **secondary endpoints** between groups in the randomised phase, treatment withdrawal was associated with a significant decline in LVEF, a significant increase in heart rate and diastolic blood pressure.* 'Что касается вторичных конечных точек между группами в рандомизированной фазе, то отмена лечения была связана со значительным снижением ФВЛЖ, значительным увеличением частоты сердечных сокращений и диастолического артериального давления'.

Коммуникативная цель семантического компонента «Нежелательные явления» – представить все нежелательные эффекты, возникшие в ходе проведения исследования: любые случаи болезни, травмы, незапланированные оперативные вмешательства и т.п., связь которых с проводимым медицинским вмешательством (профилактическим, диагностическим, лечебным или любым другим) не может быть исключена. Отсутствие нежелательных явлений тоже, как правило, отмечается автором, как и то, что их учет не проводился [4].

В тексте данный компонент вербализуется с помощью таких маркеров, как *adverse effects* 'негативные последствия', *adverse events* 'нежелательные явления', *adverse reactions* 'негативные реакции', *complications* 'осложнения' и т.п.

*Three serious **adverse events** were reported in the treatment withdrawal group: hospital admissions for urinary sepsis, non-cardiac chest pain, and an elective procedure for a pre-existing condition.* 'В группе отмены лечения было зарегистрировано три серьезных нежелательных явления: госпитализация по поводу мочевого сепсиса, несердечная боль в груди и выборная процедура по поводу ранее существовавшего заболевания'.

На основе полученных данных был разработан алгоритм поиска семантических компонентов раздела «Результаты» в научной медицинской статье на английском языке. Фрагмент алгоритма приведен на рисунке.

A1	Анализ в тексте раздела «Результаты (results)» фрагмента, расположенного непосредственно за статусом данной суперструктуры

↓

A2	Определение компонента *описание материала исследования*: фрагмент содержит термины *men, women, female, male, white, the median age/time, at enrolment* и др., а также индикаторы *We screened/enrolled/assigned/recruited, were assigned to/enrolled/included* и др.	Да →	B1

↓

A3	Определение компонента *основные результаты исследования*: фрагмент содержит маркеры *primary endpoint, primary outcome(s), primary efficacy endpoint(s), main result(s)* и др.	Да →	B2

↓

A4	Определение компонента *дополнительные результаты исследования*: фрагмент содержит языковые маркеры *secondary endpoint(s), secondary outcome(s), additional efficacy outcome(s)* и др.	Да →	B3

↓

A5	Определение компонента *нежелательные явления*: фрагмент содержит маркеры *adverse effects, adverse events, adverse reactions, complication(s), died, death* и др.	Да →	B4

↓

A6	Обработка фрагментов B1, B2, B3, B4

↓

A7	Составление реферата раздела «Результаты»

Принципиальный алгоритм поиска компонента «Результаты»
в текстах научных статей на английском языке

Следующий этап работы предусматривает сокращение отобранных фрагментов исходного текста в целях построения текста реферата.

ЛИТЕРАТУРА

1. *Залевская, А. А.* Текст и его понимание / А. А. Залевская. – Тверь : Твер. гос. ун-т, 2001. – 177 с.
2. *Меньшенина, И. А.* Структура компонента «Введение» в научных медицинских статьях на английском языке / И. А. Меньшенина // Вестн. МГЛУ. Сер. 1, Филология. – 2020. – № 4 (107). – С. 123–129.
3. *Меньшенина, И. А.* Формализация распознавания компонента «Методы» в научных медицинских статьях на английском языке / И. А. Меньшенина // Молодые ученые в инновационном поиске : сб. науч. ст. по материалам VIII Междунар. науч. конф., Минск, 29–30 мая 2019 г. / редкол.: Т. П. Карпилович (отв. ред.) [и др.]. – Минск : МГЛУ, 2020 – С. 211–216.
4. *Сайгитов, Р. Т.* Правила и рекомендации по представлению рукописей, содержащих результаты оригинальных исследований [Электронный ресурс] / Р. Т. Сайгитов // Вопросы современной педиатрии. – 2015. – № 3. – Режим доступа : https://publications.hse.ru/articles/154307819. – Дата доступа : 26.06.2020.