УДК 001.103:002
https://doi.org/10.47533/2020.1606-146X.137

Д. О. ЖАКСЫБАЕВ

Западно-Казахстанский аграрно-технический университет имени Жангир хана,
Уральск, Казахстан
darhan.03.92@mail.ru

РАЗМЕТКА КОЛЛЕКЦИИ ТЕКСТОВ КЛЮЧЕВЫМИ СЛОВАМИ – АСПЕКТЫ АВТОМАТИЗАЦИИ

В статье представлены и обсуждаются результаты автоматического индексирования ключевых слов из 27 функциональных коллекций русскоязычных текстов трех функциональных стилей: научного, публицистического и художественного. Представлен подход к обработке результатов разметки, данные о согласованности экспертов. В зависимости от характера проблем проекта, задания на проектирование предусматривают автоматизированную систему идентификации документов и ключевых слов. Цель данного исследования - выявить проблемы модифицированной системы автоматического скоринга текста (CAPT) с ключевыми словами, провести детальный анализ результатов скоринга теста, чтобы создать условия для следующего дискурса. Эти функции составляют содержание одного из этапов исследования, направленного на создание эффективного алгоритма извлечения КС для русского языка.

Ключевые слова*: разметка теста, ключевые слова, процедура отбора.*

Введение. Чтобы измерить полезность алгоритмов извлечения ключевых слов (KW), их следует протестировать на том же корпусе специализированного текста (TC), который ранее был определен KW. Для оценки эффективности применимости решений к англоязычным текстам можно выбрать из проверенных и выделенных ключевых слов TC.

Цель данного исследования – выявить проблемы модифицированной системы автоматического скоринга текста (CAPT) с ключевыми словами, провести детальный анализ результатов разметки теста, чтобы создать условия для следующего дискурса. Эти функции составляют содержание одного из этапов исследования, направленного на создание эффективного алгоритма извлечения КС для русского языка, рисунок 1.

Требования в данном случае:

Разработка и экспертиза процедур отбора;

Разработка систем использования документов для копирования;

Определять организационные и технические проблемы при написании ключевых слов и разрабатывать стратегии их решения.

Методология исследования и результаты. При организации процесса разметки текстовых корпусов важно обратить внимание на различные методы, используемые учеными, а именно: выбор состава и количества текстовых модулей, форму и содержание первичной инструкции, состав и количество экспертов, состав и методы редактирования, интерпретации и представления результатов. Поэтому в исследование[Marujo et al., 2012] было включено большое количество пометок и использовались более эвристические методы для коррекции полученных

Рисунок 1 – Этапы исследования

результатов: набор КС с содержанием шумовых слов, более длинные КС (более 10 слов), более короткое время КС (менее 30 секунд) завершали задания экспертного отбора [1, с 144].

Для оценки корпуса DUC-2001 [Wan, Xiao, 2008] два студента старших курсов были выбраны респондентами и выставили оценки по следующим правилам: количество основных линий в тексте; все противоречия должны быть разрешены путем обсуждения. В результате среднее количество строк в тексте составило 8,08, а длина слова - 2,09 слова.

В общей сложности 1020 хроник хорватских новостных статей [Mijić et al., 2010] размечался семью учеными в соответствии с полными инструкциями. Ученые обсудили ряд соглашений и слияний, но сделали это независимо друг от друга. Набор из 60 примечаний, предоставленных всеми учеными, для ознакомления с уровнем соответствия между комментаторами. Оставшиеся 960 рукописей были разделены на 120 рукописей и описаны каждым из семи ученых. Основные положения трех экспертов, отобранных на этапе объединения, считались золотым стандартом.

Четыре раздела интервью на голландском языке (по 1000-2000 статей в каждом) были написаны 357 работающими участниками за денежное вознаграждение [Sterckx & s., 2018]. 26% статей были написаны восемью-десятью академиками. Каждый участник получил образец документа с инструкциями, видеоматериалы о процессе оценки и обновленные ПС.

Автоматизированное и аннотирование контента ключевыми словами необходимо для целенаправленных действий обучающихся, менеджеров и программно-аппаратных сообществ, ориентированных на разработку метаинформации – CS-описания для каждого элемента, рис. 2.

Ход разметки состоит из трех этапов. На первом этапе оценки CS выставляются напрямую, эксперты взаимодействуют с текстом в определенный момент времени, а в качестве альтернативы они основываются на оригинальной работе (инструкции), которая присваивает ключевые слова исходному тексту.

В процессе редактирования вычисляются промежуточные числа для операции тегирования, эксперты ранжируются по надежности и силе, и результат фиксируется: в документах CW, исключены дубликаты и ошибки в основных строках, ключевые фразы нормализуются, унифицируются и ранжируются [2, с 39].

Рисунок 2 – Эффективный и интуитивно понятный способ регистрации разметки текста

Методы управления метками включают активное управление (обновление) метками (закон конфигурации и пользовательский опыт, параметры для распределения в базе данных) и планирование (остановка планирования проекта, монтаж конфигурации).

Работа CAPT основана на архитектуре клиент-сервер с использованием сервера IIS. Часть сервера размещена на C# в связи с сервером баз данных MS SQL, клиентская часть управляется технологией ASP.NET Razor, JavaScript и библиотекой jQuery.

В первом исследовании CAPT были проведены маломасштабные эксперименты по регрессии КС. Специалистам были предоставлены три проекта (определенные для использования сервиса), каждый из которых состоял из девяти шагов с тремя практическими этапами «Худломер»:

1) исторический, художественный и научный. В то же время различия между проектами заключаются в разных методах описания КС:

2. Введение KS вручную, используя любое слово или фразу без знаков препинания;

3. Комбинирование первого и второго методов. Пользователи могли выполнить все три задания по порядку, проигнорировать предыдущие тексты и вернуться обратно. Каждое задание начиналось с краткого описания, в котором намеренно опускалось слово «ключевое слово», его описание и стиль. Вместо этого ему предписывается выбирать слова (фразы) из описательного текста в контексте [3, с 555].

Для участия в исследовании мы использовали 24 пункта (включая работающих писателей) – русскоговорящих, в том числе 19 студентов 3-5 курсов университета, пять человек с высшим образованием. Исследование длилось 18 дней, а общая продолжительность программы пользователя CAPT составила 5 дней 12 часов.

В исследовательских целях на модельном тексте был рассчитан ряд количественных показателей – прежде всего, сила и энтропия частотного словаря, отражающие богатство художественного текста [Романишин, 2016], Маковецкая сложность, сте-

пень сжатия. Энтропия частотного словаря рассматривается как старая энтропия по всем элементам словаря:

$$H = \sum_{i=1}^{n} pH = -\sum_{i-1}^{n} p_i \log p_i \quad (1)$$

Здесь pi – это умножение элемента e частотного словаря.

Частота частотных словарей рассчитывается по первым десяти элементам первых частотных элементов частотного словаря:

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_1 - w_n)}} \quad (2)$$

Где P - сумма объектов в словаре, N - количество объектов (N = 10).

Исторические и научные тексты о сложности и энтропии частотного словаря неизбежны, рис. 3.

Полученные результаты можно объяснить ошибками в составлении текста модели маркера. Таким образом, научная и литературная система была предвзятым экономическим дискурсом, а журналистика – экспериментальным экспериментом.

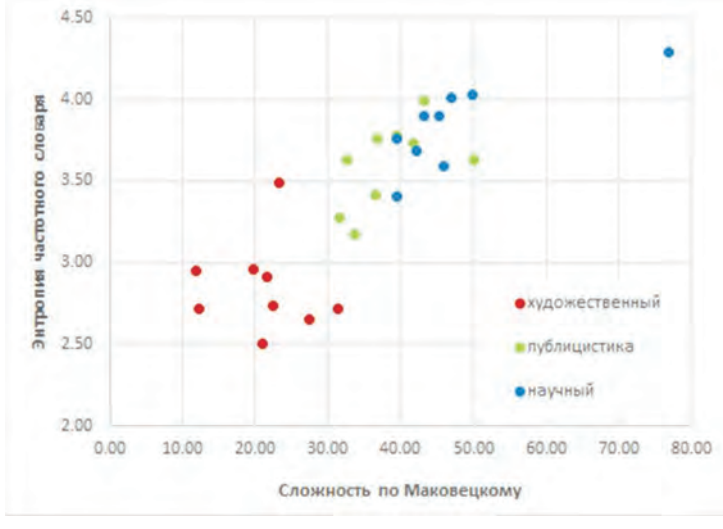

Рисунок 3 – Описание текста экспериментальной модели на диаграмме энтропия-сложность.

Предварительная обработка списка ключевых слов

В редактирование были включены речи, дублирующие термины (размер словаря 264 слова), ошибки исправления, найденные CW с помощью библиотеки Hunspell1 со встроенными словарями русского языка. Числа, рассчитанные по длине основных строк в тексте, затем использовались для выполнения заданий, количества КС для каждого слова.

Для того чтобы отфильтровать результаты маркировки «мусорных» систем, против каждого эксперта были использованы следующие показатели, рисунок 4 [4, с 399].

1. Сотни имен CS были исключены из оценки из-за двойной скорости чтения (зеленый - нет закономерности, коричневый – менее 10%, красный – более 10%);

2. средняя скорость маркировки слов в секунду (x-чтение);

3. Коэффициент вариации скорости разметки;

4. Среднее количество основных линий в тексте (порядковое чтение);

5. коэффициент разности сигналов предыдущего показателя (размер маркера).

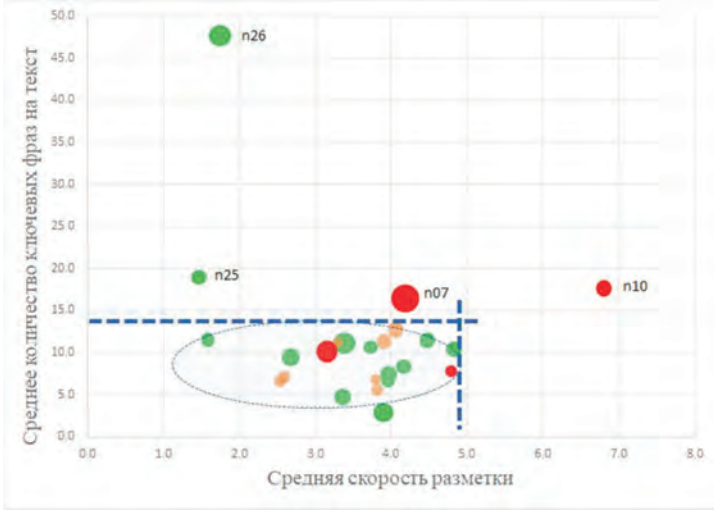

Рисунок 4 – Представление пользователей в соответствии с рядом критериев

Из рисунка видно, что для получения объективной информации необходимо учитывать уровень имеющихся навыков измерений и расчетов. Например, пользователь n26 значительно отличается по длине основных линий, но другие параметры находятся в нормальном диапазоне. Поэтому, помимо скорости и качества запасов, эксперты также сравнили тип показателей импортных КК:

1. средняя длина основной строки всего текста в словах ;

2. коэффициент вариации средней длины магистрали для всех записей;

3. Часть информации об остановке и стоп-сигналы, определенные в выбранном CW.

Кроме того, были созданы коллекции экспертов. Мы использовали два метода из библиотеки SCIKIT Python, которая стала стандартом де-факто в машинном обучении. Метод MeanShift использует алгоритм, основанный на центре, который вычисляет их по расположению в заданном пространстве. Нет необходимости определять количество компонентов, получаемых в результате применения этого метода. Метод Kmeans – это алгоритм кластеризации, который уменьшает важность инерции или разницы углов. Для этого алгоритма необходимо указать количество выходных узлов. Исследования показали, что все алгоритмы дают практически одинаковые результаты, а все эксперты делятся на 3 области, рисунок 5.

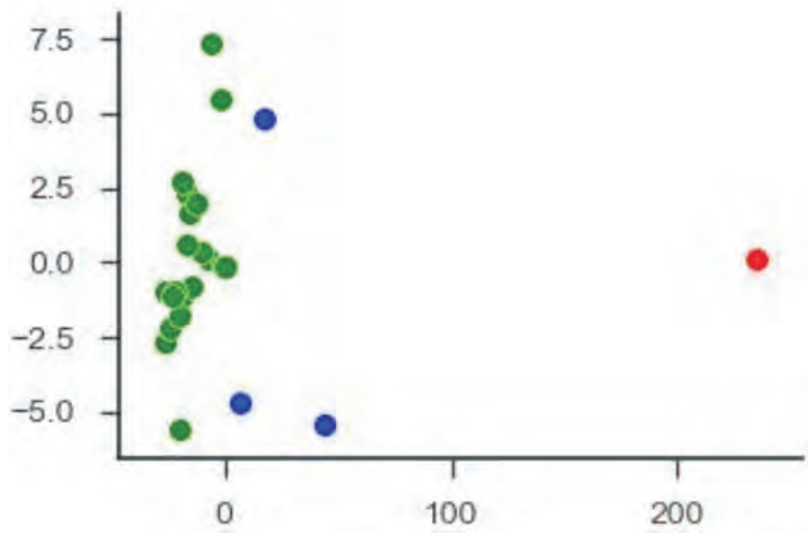

Рисунок 5 – Результаты экспертов по пулингу в двух основных категориях

Группировка проводилась на исходных признаках без значительной предварительной обработки. Для наглядности использовался метод больших секций, и все компоненты рассчитывались в первых двух секциях. Следует отметить, что использование комбинации после применения метода главных компонент дало схожие результаты [5, с 60].

Основные положения фильтрации таковы: в отличие от больших кластеров меньшие кластеры на самом деле должны быть удалены из «мусорных» систем в первую очередь. Например, на рисунке 4 пользователь отмечен красным цветом. В этом случае внешний вид можно исправить, посмотрев и проанализировав его цифры. Поэтому при наличии большого количества экспертов и увеличении числа тестируемых статистических методов провести визуальный анализ не удастся. По этой причине планируется использовать кластерный анализ для осуществления автоматической фильтрации.

По результатам опроса результаты четырех участников были исключены без дальнейшего рассмотрения.

Для того чтобы использовать данные участников проекта (баллы CS), уровень согласия экспертов должен быть указан в утвержденных пределах, установленных конкретным исследованием, т.е. быть объективным [6, с 503].

Степень консенсуса среди экспертов может быть измерена различными способами [Artstein, Poesio, 2009], причем наиболее известный компонент - это % элементов,который предпочитают многие ученые. Средний показатель нестабильности ПС составил 6,7%, что является слабым показателем. Отметим, что проблема незначительных пробелов в выборе символов CS также важна для больших коллекций документов (более 1000 экземпляров) [Sterckx & s., 2018]. Гистограмма количества совпадений COP, выделенных из записей, показана на рисунке 6. Строки, написанные тем же пользователем, достигли 2056 элементов и не отображаются.

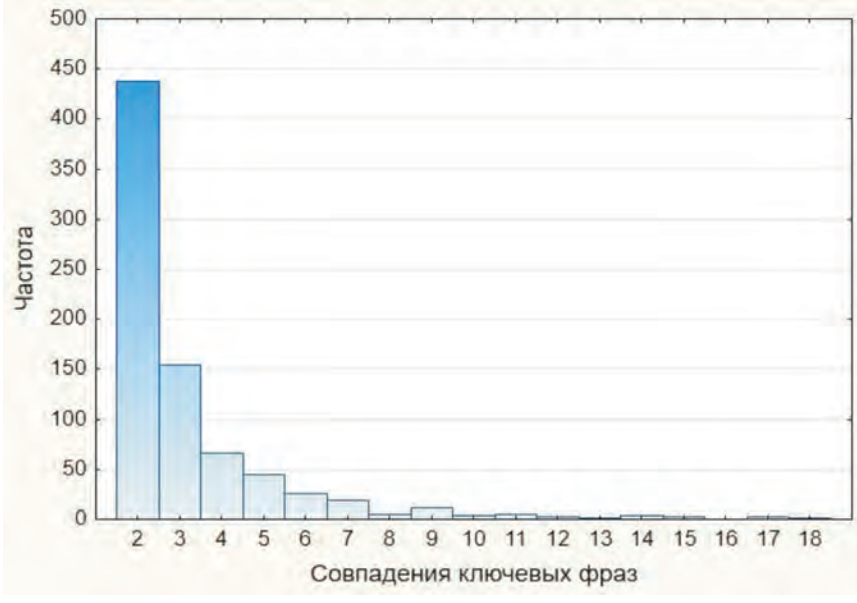

Рисунок 6 – Подбор ключевых слов

Результаты и выводы. По результатам анализа можно выделить следующие показатели «хорошего» эксперта:

1. Стабильность и «вдумчивость» при чтении;

2. Размер описательного КС составляет 5-15 пунктов, а длина предложения не превышает пяти слов;

3. Предложения в описании КС устойчивы, отсутствуют знаки препинания (конверты, местоимения, предлоги).

В ходе исследования в качестве промежуточного результата были выявлены технические и организационные недостатки, и можно было разработать руководство для конкретной задачи: [7, с 855].

1) Экспертам необходимо дать организованную работу (формализованные задания), чтобы показать примеры.

2) Экспертам рекомендуется представить, как минимум, три учебных материала и первичного ранжирования, а также устранить результаты обучения (отсутствующие в существующем наборе данных);

3) Учитывать роль пользователя при рассмотрении времени, необходимого для маркировки;

4) Отображение слов из нетекстовых текстов, что кажется очень эргономичным и удобным для пользователей;

5) Трудно планировать работу эксперта без определенных финансовых ресурсов. Необходимо улучшить подход участника к мотивации и проинформировать провайдера о его текущем статусе (уровне уверенности);

6) По возможности, эксперты должны быть тщательно отобраны по количеству профессиональных групп, половому признаку, возраст и социальный статус;

7) Нужен инструмент для управления процессом рейтингования в режиме реального времени.

8) Оценка и продвинутые навыки высоко развиты, но вопрос о границах арифметического значения и профессиональных способностей требует отдельного изучения.

С учетом этого и в целях продолжения двуязычной работы готовится следующая версия SART. Предлагаются следующие функциональные изменения, рис. 9:

1. Обновление пользовательского интерфейса (в настоящее время доступен на русском и английском языках);

2. Установить ограничение по времени для мониторинга активности пользователя;

3. Создание регулярных выражений для проверки выбранных/назначенных слов;

4. Считывайте различную информацию о ходе работы маркера в режиме реального времени.

Применение технологии также было переработано - серверный компонент был перенесен на основной фреймворк .net, а клиентский компонент - на фреймворк Webix1. Требуется другая интерпретация речи.

ЛИТЕРАТУРА

1 Вормсбехер В.Ф. 100 страниц в час / В.Ф Ворм- сбехер, В.А Кабин. – Кемерово: Кемеровское книжное издательство, 1980. – 144 с. [Vormsbeher V.F. 100 stranic v chas / V.F Vormsbeher, V.A Kabin. – Kemerovo: Kemerovskoe knizhnoe izdatel'stvo, 1980. – 144 c.]

2 Романишин Г.В. Исследование лексического богатства научных текстов / Г.В. Романишин // Новые информационные технологии в автоматизированных системах. – 2016. – №. 19 – С. 39- 42. [Romanishin G.V. Issledovanie leksicheskogo bogatstva nauchnyh tekstov / G.V. Romanishin // Novye informacionnye tehnologii v avtomatizirovannyh sistemah. – 2016. – №. 19 – S. 39- 42.]

3 Artstein R. Survey Article: Inter-Coder Agreement for Computational Linguistics / R. Artstein, M Poesio // Computational Linguistics. – 2009. – Vol. 34. – Iss 4. – pp. 555–596.

4 Marujo L. Supervised Topical Key Phrase Extrac- tion of News Stories using Crowdsourcing / L. Marujo, A. Gershman, J.G. Carbonell, R.E Frederk- ing, J.P. Neto // 8th International Conference on Language Resources and Evaluation (LREC 2012). – 2012. – pp. 399-403.

5 Mijić J. Robust Keyphrase Extraction For A Large- Scale Croatian News Production System / J. Mijić, J. Šnajder, B. Dalbelo Bašić // Proceedings of the Seventh International Conference on Formal Ap- proaches to South Slavic and Balkan Languages. – 2010. – pp. 59–66.

6 Sterckx L. Creation and evaluation of large keyphrase extraction collections with multiple opinions / L. Sterckx, T. Demeester, J. Deleu and C. Develder // Language Resources and Evaluation. – 2018. – Vol. 52. – Iss 2. – pp. 503–532.

7 Wan X. Single Document Keyphrase Extraction Us- ing Neighborhood Knowledge / X. Wan, J. Xiao // Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. – 2008. – pp 855-860.

REFERENCES

1 Vormsbekher V.F. 100 stranic v chas / V.F Vorm- sbekher, V.A Kabin. – Kemerovo: Kemerovskoe knizhnoe izdatel'stvo, 1980. – 144 c. [Vormsbeher V.F. 100 stranic v chas / V.F Vorm- sbeher, V.A Kabin. – Kemerovo: Kemerovskoe knizhnoe izdatel'stvo, 1980. – 144 c.]

2 Romanishin G.V. Issledovanie leksicheskogo bogatstva nauchnyh tekstov / G.V. Romanishin // Novye informacionnye tekhnologii v avtomatizirovannyh sistemah. – 2016. – №. 19 – S. 39-

42. [Romanishin G.V. Issledovanie leksicheskogo bogatstva nauchnyh tekstov / G.V. Romanishin // Novye informacionnye tehnologii v avtomatizirovannyh sistemah. – 2016. – №. 19 – S. 39- 42.]

3 Artstein R. Survey Article: Inter-Coder Agreement for Computational Linguistics / R. Artstein, M Poesio // Computational Linguistics. – 2009. – Vol. 34. – Iss 4. – pp. 555–596.

4 Marujo L. Supervised Topical Key Phrase Extrac- tion of News Stories using Crowdsourcing / L. Marujo, A. Gershman, J.G. Carbonell, R.E Frederk- ing, J.P. Neto // 8th International Conference on Language Resources and Evaluation (LREC 2012). – 2012. – pp. 399-403.

5 Mijić J. Robust Keyphrase Extraction For A Large- Scale Croatian News Production System / J. Mijić, J. Šnajder, B. Dalbelo Bašić // Proceedings of the Seventh International Conference on Formal Ap- proaches to South Slavic and Balkan Languages. – 2010. – pp. 59–66.

6 Sterckx L. Creation and evaluation of large keyphrase extraction collections with multiple opinions / L. Sterckx, T. Demeester, J. Deleu and C. Develder // Language Resources and Evaluation. – 2018. – Vol. 52. – Iss 2. – pp. 503–532.

7 Wan X. Single Document Keyphrase Extraction Us- ing Neighborhood Knowledge / X. Wan, J. Xiao // Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. – 2008. – pp 855-860.

D. O. ZHAXYBAYEV

Zhangir khan West Kazakhstan Agrarian Technical University,
Uralsk, Kazakhstan
darhan.03.92@mail.ru

MARKING A COLLECTION OF TEXTS WITH THE KEYWORDS – AUTOMATION ASPECTS

This article presents and discusses the results of automatic indexing of keywords in 27 functional collections of Russian texts in three functional styles: scholarly, journalistic and fiction. The approach to the processing of markup results is presented and the data on the consistency of experts are given. Depending on the nature of the project's problems, the design tasks provide for an automated system of document and keyword identification. The aim of this study is to identify the problems of a modified automatic text scoring system (CAPT) with keywords and to analyse in detail the results of the scoring test in order to create the conditions for the next discourse. These functions constitute the content of one of the research stages aimed at creating an effective algorithm for CS extraction for Russian language.

Keywords*: test markup, keywords, selection procedure*

Д. О. ЖАКСЫБАЕВ

Жәңгір хан атындағы Батыс Қазақстан аграрлық-техникалық университеті,
Орал, Қазақстан
darhan.03.92@mail.ru

МӘТІНДЕР ЖИНАҒЫН КІЛТ СӨЗДЕРМЕН БЕЛГІЛЕУ – АВТОМАТТАНДЫРУ АСПЕКТІЛЕРІ

Мақалада үш функционалды стильдегі орыс тіліндегі мәтіндердің 27 функционалды топтамасынан кілт сөздерді автоматты индекстеу нәтижелері ұсынылған және талқыланады: ғылыми,

публицистикалық және көркем. Таңбалау нәтижелерін өңдеу тәсілі, сарапшылардың келісімділігі туралы мәліметтер ұсынылған. Жоба мәселелерінің сипатына байланысты жобалау тапсырмалары құжаттар мен кілт сөздерді сәйкестендірудің автоматтандырылған жүйесін қарастырады. Бұл зерттеудің мақсаты-өзгертілген мәтінді автоматты скоринг жүйесінің (CAPT) мәселелерін кілт сөздермен анықтау, келесі дискурсқа жағдай жасау үшін тест скорингінің нәтижелеріне егжей-тегжейлі талдау жүргізу. Бұл функциялар орыс тілі үшін КС алудың тиімді алгоритмін құруға бағытталған зерттеу кезеңдерінің бірінің мазмұнын құрайды.

***Түйін сөздер**: тестті белгілеу, кілт сөздер, іріктеу процедурасы.*