**Результаты по табличному бенчмарку RF-200**

**Статистика по размеченному набору табличных данных RF-200 (ru-facts-200):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Предметная область** | **Кол-во таблиц** | **Кол-во столбцов** | **Кол-во ячеек** | **Кол-во пустых ячеек** |
| Локации | 33 | 153 | 1992 | 99 |
| Спорт | 26 | 188 | 4786 | 614 |
| Киноиндустрия и театры | 16 | 74 | 1464 | 44 |
| Политика | 15 | 68 | 915 | 81 |
| История | 9 | 33 | 877 | 156 |
| Кинонаграды | 8 | 42 | 588 | 3 |
| Автоспорт | 8 | 45 | 431 | 10 |
| Архитектурные сооружения | 8 | 47 | 753 | 124 |
| Печатные издания | 7 | 29 | 554 | 48 |
| Музыка | 6 | 21 | 618 | 6 |
| Измерения | 6 | 23 | 248 | 11 |
| Праздники и мероприятия | 6 | 20 | 263 | 12 |
| Природные объекты | 6 | 26 | 359 | 21 |
| Авиация | 5 | 28 | 499 | 109 |
| Медиа | 5 | 16 | 221 | 17 |
| Организации и объединения | 5 | 37 | 640 | 55 |
| Продукты питания | 4 | 14 | 194 | 0 |
| Астрономия | 4 | 16 | 800 | 0 |
| Национальности и этносы | 4 | 12 | 243 | 10 |
| Религия | 4 | 11 | 102 | 0 |
| Торговля и финансы | 3 | 12 | 280 | 0 |
| Награды и премии | 3 | 13 | 342 | 52 |
| Реслинг | 3 | 14 | 472 | 41 |
| Компьютерные игры | 3 | 14 | 906 | 143 |
| Энергетика | 2 | 16 | 880 | 63 |
| Телешоу | 1 | 7 | 63 | 0 |
| **ИТОГО** | **200** | **979** | **19490** | **1719** |

* Среднее количество столбцов на 1 таблицусоставило: **4,89**
* Среднее количество ячеек на 1 таблицусоставило: **97,45**
* Доля пустых ячеек в **RF-200** составила: **8,82%**

**Статистика по модели предметной области (используемое количество элементов):**

|  |  |
| --- | --- |
| **Элемент** | **Кол-во** |
| Тип концепта | 29 |
| Тип значения характеристики | 9 |
| Тип характеристики концепта | 281 |
| Тип связи | 50 |
| Тип характеристики связи | 22 |
| **ИТОГО** | **391** |

**Итоговая экспериментальная оценка на RF-200:**

(!) Лучшие значения оценки в таблице отмечены зеленым цветом.

**CA – Семантический Анализатор**

**АОТ – Автоматический Обработчик Таблиц (привязанный к результатам СА)**

**АОТ\* – Автоматический Обработчик Таблиц (отвязан от СА)**

**P – Precision / R – Recall**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Задача** | **СА** | | | **АОТ** | | | **АОТ\*** | | |
| **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| NERC | 0,663 | 0,535 | 0,547 | 0,646 | 0,614 | 0,601 | 0,570 | 0,771 | 0,633 |
| RELEXT | 0,000 | 0,000 | 0,000 | 0,347 | 0,252 | 0,277 | 0,405 | 0,334 | 0,347 |
| NERC + RELEXT | **0,332** | **0,268** | **0,274** | **0,497** | **0,433** | **0,439** | **0,488** | **0,552** | **0,490** |

**Экспериментальная оценка по отдельным предметным областям для NERC:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Предметная область** | **СА** | | | **АОТ** | | | **АОТ\*** | | |
| **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Локации | 0,706 | 0,610 | 0,591 | 0,715 | 0,748 | 0,708 | 0,608 | 0,841 | 0,691 |
| Спорт | 0,580 | 0,678 | 0,601 | 0,547 | 0,699 | 0,590 | 0,406 | 0,749 | 0,490 |
| Киноиндустрия и театры | 0,669 | 0,481 | 0,501 | 0,561 | 0,576 | 0,551 | 0,510 | 0,842 | 0,611 |
| Политика | 0,718 | 0,727 | 0,708 | 0,853 | 0,808 | 0,812 | 0,643 | 0,862 | 0,719 |
| Кинонаграды | 0,569 | 0,389 | 0,415 | 0,406 | 0,514 | 0,428 | 0,163 | 0,579 | 0,244 |
| История | 0,734 | 0,691 | 0,699 | 0,677 | 0,691 | 0,673 | 0,713 | 0,771 | 0,723 |
| Музыка | 0,844 | 0,632 | 0,621 | 0,827 | 0,792 | 0,751 | 0,526 | 0,859 | 0,623 |
| Автоспорт | 0,676 | 0,681 | 0,657 | 0,684 | 0,748 | 0,700 | 0,536 | 0,801 | 0,616 |
| Архитектурные сооружения | 0,527 | 0,450 | 0,462 | 0,565 | 0,579 | 0,513 | 0,444 | 0,742 | 0,532 |
| Торговля и финансы | 0,724 | 0,309 | 0,414 | 0,667 | 0,467 | 0,524 | 0,648 | 0,867 | 0,729 |
| Энергетика | 0,749 | 0,546 | 0,607 | 0,831 | 0,546 | 0,647 | 0,572 | 0,679 | 0,616 |
| Печатные издания | 0,647 | 0,662 | 0,638 | 0,592 | 0,728 | 0,642 | 0,602 | 0,770 | 0,669 |
| Измерения | 0,187 | 0,167 | 0,167 | 0,134 | 0,186 | 0,147 | 0,408 | 0,491 | 0,435 |
| Праздники и мероприятия | 0,625 | 0,782 | 0,688 | 0,649 | 0,798 | 0,707 | 0,627 | 0,798 | 0,693 |
| Природные объекты | 0,819 | 0,506 | 0,555 | 0,893 | 0,675 | 0,737 | 0,742 | 0,944 | 0,817 |
| Авиация | 0,485 | 0,581 | 0,499 | 0,516 | 0,696 | 0,580 | 0,511 | 0,722 | 0,590 |
| Медиа | 0,703 | 0,737 | 0,711 | 0,768 | 0,742 | 0,754 | 0,799 | 0,835 | 0,816 |
| Организации и объединения | 0,570 | 0,617 | 0,562 | 0,655 | 0,779 | 0,699 | 0,461 | 0,779 | 0,517 |
| Продукты питания | 0,777 | 0,255 | 0,378 | 0,583 | 0,312 | 0,383 | 0,579 | 0,583 | 0,579 |
| Астрономия | 0,571 | 0,159 | 0,228 | 0,547 | 0,231 | 0,305 | 0,647 | 0,807 | 0,701 |
| Национальности и этносы | 0,950 | 0,854 | 0,883 | 0,998 | 0,937 | 0,963 | 0,998 | 1,000 | 0,999 |
| Религия | 0,502 | 0,406 | 0,426 | 0,567 | 0,505 | 0,520 | 0,765 | 0,905 | 0,822 |
| Телешоу | 0,455 | 0,234 | 0,309 | 0,469 | 0,359 | 0,407 | 0,397 | 0,781 | 0,526 |
| Награды и премии | 0,873 | 0,373 | 0,491 | 0,642 | 0,389 | 0,481 | 0,371 | 0,444 | 0,403 |
| Реслинг | 0,764 | 0,794 | 0,768 | 0,738 | 0,728 | 0,716 | 0,698 | 0,767 | 0,704 |
| Компьютерные игры | 0,818 | 0,595 | 0,660 | 0,720 | 0,727 | 0,693 | 0,454 | 0,819 | 0,583 |
| **ИТОГО** | **0,663** | **0,535** | **0,547** | **0,646** | **0,614** | **0,601** | **0,570** | **0,771** | **0,633** |

**Экспериментальная оценка по отдельным предметным областям для RELEXT:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Предметная область** | **СА** | | | **АОТ** | | | **АОТ\*** | | |
| **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| Локации | 0 | 0 | 0 | 0,279 | 0,312 | 0,285 | 0,277 | 0,333 | 0,286 |
| Спорт | 0 | 0 | 0 | 0,335 | 0,205 | 0,230 | 0,300 | 0,228 | 0,226 |
| Киноиндустрия и театры | 0 | 0 | 0 | 0,513 | 0,248 | 0,321 | 0,271 | 0,217 | 0,230 |
| Политика | 0 | 0 | 0 | 0,461 | 0,380 | 0,407 | 0,427 | 0,443 | 0,421 |
| Кинонаграды | 0 | 0 | 0 | 0,000 | 0,000 | 0,000 | 0,006 | 0,049 | 0,011 |
| История | 0 | 0 | 0 | 0,398 | 0,313 | 0,341 | 0,367 | 0,313 | 0,332 |
| Музыка | 0 | 0 | 0 | 0,198 | 0,350 | 0,252 | 0,331 | 0,447 | 0,371 |
| Автоспорт | 0 | 0 | 0 | 0,448 | 0,266 | 0,304 | 0,286 | 0,235 | 0,245 |
| Архитектурные сооружения | 0 | 0 | 0 | 0,251 | 0,165 | 0,193 | 0,429 | 0,336 | 0,357 |
| Торговля и финансы | 0 | 0 | 0 | 0,111 | 0,111 | 0,111 | 0,111 | 0,111 | 0,111 |
| Энергетика | 0 | 0 | 0 | 0,500 | 0,250 | 0,333 | 0,500 | 0,250 | 0,333 |
| Печатные издания | 0 | 0 | 0 | 0,410 | 0,294 | 0,328 | 0,417 | 0,368 | 0,375 |
| Измерения | 0 | 0 | 0 | 0,000 | 0,000 | 0,000 | 0,154 | 0,134 | 0,147 |
| Праздники и мероприятия | 0 | 0 | 0 | 0,833 | 0,526 | 0,609 | 0,833 | 0,526 | 0,609 |
| Природные объекты | 0 | 0 | 0 | 0,197 | 0,206 | 0,201 | 0,264 | 0,279 | 0,271 |
| Авиация | 0 | 0 | 0 | 0,525 | 0,297 | 0,360 | 0,529 | 0,302 | 0,365 |
| Медиа | 0 | 0 | 0 | 0,376 | 0,271 | 0,302 | 0,965 | 0,555 | 0,684 |
| Организации и объединения | 0 | 0 | 0 | 0,784 | 0,438 | 0,534 | 0,717 | 0,438 | 0,515 |
| Продукты питания | 0 | 0 | 0 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |
| Астрономия | 0 | 0 | 0 | 0,000 | 0,000 | 0,000 | 0,558 | 0,591 | 0,568 |
| Национальности и этносы | 0 | 0 | 0 | 0,750 | 0,628 | 0,669 | 1,000 | 0,878 | 0,919 |
| Религия | 0 | 0 | 0 | 0,312 | 0,328 | 0,319 | 0,687 | 0,703 | 0,694 |
| Телешоу | 0 | 0 | 0 | 0,000 | 0,000 | 0,000 | 0,043 | 0,079 | 0,056 |
| Награды и премии | 0 | 0 | 0 | 0,308 | 0,154 | 0,205 | 0,308 | 0,154 | 0,205 |
| Реслинг | 0 | 0 | 0 | 0,427 | 0,430 | 0,422 | 0,287 | 0,333 | 0,308 |
| Компьютерные игры | 0 | 0 | 0 | 0,627 | 0,378 | 0,466 | 0,471 | 0,379 | 0,388 |
| **ИТОГО** | **0** | **0** | **0** | **0,347** | **0,252** | **0,277** | **0,405** | **0,334** | **0,347** |

**Выводы по полученной оценки производительности:**

* Оценка производилась отдельно по каждой таблице с использованием только определенной часть модели предметной области (подмножества типов).
* Точность NERC оказалась немного выше для семантического анализатора. Это связано с тем, что семантический анализатор может точно выделять необходимые значения в ячейках. В то время как автоматический обработчик таблиц всегда выделяет значения ячеек целиком.
* Полнота NERC оказалась выше для автоматического обработчика таблиц за счет того, что семантический анализатор может пропускать некоторые значения ячеек в столбце. Особенно если они принадлежат к редким NERC-меткам (например, каким-то редко встречаемым событиям, мероприятиям и т.п.). В то время как автоматический обработчик таблиц всегда выделяет все значения ячеек в столбце.
* Оценки RELEXT для семантического анализатора оказались нулевыми из-за того, что он выделяет связи и характеристики только внутри одной ноды (ячейки). В то время как автоматический обработчик таблиц может выделять связи и характеристики между разными нодами (ячейками).
* В целом итоговые оценки NERC+RELEXT для автоматического обработчика таблиц оказалась ожидаемо выше, чем для семантического анализатора за счет своей направленности на обработку таблиц.

**Ошибки, опечатки и особенности в собранном наборе табличных данных:**

* Мусорные теги HTML в некоторых значениях ячеек.
* Разные единицы измерения, которые могут приводить к разным разновидностям характеристик. Пример: «*локации*» – в 6 таблице измерение численности населения в млн. чел., а в таблице 7 в тыс. чел. Соответственно, у концепта могут быть две характеристики: «*население в млн. чел.*» и «*население в тыс. чел.*».

**Причины (проблемы), повлиявшие на низкую оценку производительности:**

* Работа автоматического обработчика таблиц полностью основана на результатах работы семантического анализатора, т.е. если семантический анализатор в столбце ничего не выделил (не нашел никакие NERC-метки), то в этом столбце автоматический обработчик таблиц также ничего не будет искать.
* Автоматический обработчик таблиц не извлекает характеристики связей.
* В ячейках с идентифицирующими характеристиками могут попадаться пустые ячейки или может стоять прочерк или «н/д». Пример: «*кинонаграды*» – в 8 таблице.
* В одном столбце представлены разные концепты с характеристиками. Примеры: «*кинонаграды*» – в 10 таблице (персона и фильм); «*политические партии*» – в 1 таблице (партии); «*локации*» – в 10 таблице (государство и организация).
* В одной ячейке представлены разные концепты или характеристики. Примеры: «*музыка*» – в 2 таблице (столбец «чарт» в котором есть название чарта и его местонахождение); «*музыка*» – в 5 таблице (столбец «детали» в котором есть дата, организация и формат); «*торговля и финансы*» – в 1 таблице (первый столбец содержит страны и регионы/континенты).
* В одной ячейке представлены множественные значения концептов или характеристик одного типа. Примеры: «*киноиндустрия и театры*» – в 7 таблице; «*энергетика*» – в 6 таблице; «*телешоу*» – в 1 и 2 таблице.
* Концепт с его идентифицирующей характеристикой расположен вне таблицы, а в самой таблице есть только характеристики этого концепта. Примеры: «*музыка*» – в 1 таблице (второй столбец с оценкой) или в 8 таблице, где столбец «дата» и «формат» относятся к концепту песни; «автоспорт» – в 5 таблице (столбец «СУ» указывает на характеристику связи пилота с мероприятием); «*архитектурные сооружения*» – в 3 таблице (все столбцы относятся к дому или ЖК); «*торговля и финансы*» – с 2 по 5 таблицу; «*энергетика*» – в 4 таблице; «*организации и объединения*» – в 4 и 5 таблице; «*телешоу*» – в 2 и 3 таблице; «*награды и премии*» – в 2 и 3 таблице.
* Отобранная таблица является матричной или можно говорить, что одна таблица содержит несколько подтаблиц. Примеры: «*политика*» – 1 и 4 таблица; «*энергетика*» – в 1 и 2 таблице.
* Названия заголовков являются названиями характеристик концепта. Примеры: «*музыка*» – в 5 и 6 таблице; «*спорт*» – в 11 и 16 таблице; «*политика*» – в 16 таблице.
* Названия заголовков являются названиями характеристик связи. Примеры: «*спорт*» – в 2 и 7 таблице; «*кинонаграды*» – в 3 таблице.
* Характеристика концепта или связи является составной и распределена в нескольких столбцах. Пример: «*спорт*» – в 28 таблице характеристика связи «Счет» представлена в трех столбцах.
* Характеристика концепта или связи является составной и распределена в одной ячейке. Примеры: «*локации*» – в 32 таблице есть составная характеристика концепта (старые название); «*киноиндустрия и театры*» – в 18 таблице присутствуют составные типы характеристик (даты выхода фильма и страны); «*печатные издания*» – в 7 таблице есть составная характеристика концепта (год, состоящий из диапазона); «*компьютерные игры*» – в 3 таблице есть составная характеристика концепта (карьера, состоящая из дат).
* Не понятна семантика столбцов. Примеры: «*автоспорт*» – в 3 таблице в столбце «Причина»; «*архитектурные сооружения*» – в 8 таблице два последних столбца; «*спорт*» – в 28 таблице; «*локации*» – в 31 таблице.
* Наличие вычисляемых значений ячеек в столбцах. Пример: «*автоспорт*» – в 6 таблице в столбце время.

(!) Предметная область «*Национальности и этносы*» – является сложной областью, т.к. похоже, там есть концепт типа «*Перепись населения*» у которой нет именной идентифицирующей характеристики, а есть комбинация идентифицирующих характеристик типа даты и страны, где проводилась эта перепись. Также в заголовках содержится информация о датах и странах + есть поля, которые не являются национальностью!

**Что требуется доработать исходя из выявленных причин (проблем):**

* Добавить возможность транспонирования таблиц в ручной обработчик таблиц (*новая задача*).
* Отвязать работу автоматического обработчика таблиц от семантического анализатора, т.е. добавить в обработку столбцы, где семантический анализатор ничего не выделил (*новая задача*).
* Реализация сборки модели предметной области на основе тестового стенда для тестирования автоматического обработчика таблиц на всей модели предметной области целиком. В соответствии с этим необходимо будет пересчитать оценку производительности (*задача TALIE-932*).
* Добавить корректную обработку извлечения идентифицирующих характеристик с учетом пустых ячеек, прочерков или специальных символов, например, «н/д» (*новая задача*).
* Добавить извлечение характеристик связей из таблиц в автоматический обработчик таблиц (*задача TALIE-626*).
* Добавить извлечение фактов составных характеристик из таблиц в автоматический обработчик таблиц. При этом составная характеристика может собираться как внутри одной ячейки, так и быть собрана из ячеек разных столбцов (*задача TALIE-757*).
* Добавить обработку множественных однотипных значений (концептов или характеристик) в ячейках (*новая задача*).
* Добавить извлечение фактов из таблиц с использованием внешнего контекста таблицы. В частности, связь с фактами, которые расположены в остальном документе, например в названии таблицы (*задача TALIE-755*).