# Алгоритм автоматического выделения жалоб пациентов из историй болезни

**Клышинский Э.С.**, НИУ Высшая школа экономики, МИЭМ
eklyshinsky@hse.ru
**Грибова В.В.**, Институт автоматики и процесса управления ДВО РАН
**Шахгельдян К.И.**, Владивостокский государственный университет экономики и сервиса, Дальневосточный федеральный университет
carinashakh@gmail.com
**Шалфеева Е.А., Окунь Д.Б.**, Институт автоматики и процесса управления ДВО РАН
**Горбач Т.А.**, Дальневосточный федеральный университет
**Карпик О.В.**, Институт прикладной математики им. М.В. Келдыша РАН

**Аннотация**

В настоящее время медицинские организации накапливают большой объем неструктурированной информации о пациентах, для обработки которой требуются алгоритмы формализации текста. Примером такой задачи является автоматическое извлечение жалоб пациентов и их характеристик из текстов историй болезни. В данной работе предлагается алгоритм, использующий синтаксический анализ текста истории болезни, с дальнейшим уточнением семантики при помощи онтологии, содержащей описание жалоб в формализованном виде. Алгоритм апробирован на данных 3 тыс. историй болезни отделения нейрохирургии.

## 1 Введение

Создание медицинских информационных систем позволило начать массовое накопление информации об историях болезни пациентов. Ранее подобное накопление велось в гораздо меньших масштабах, являясь скорее инициативой отдельных врачей или базой для проведения частных исследований.

Накопленная информация нуждается в разработке масштабируемых решений для ее обработки. Работы в этом направлении были начаты еще в конце XX века. Так, в 1986 г. Национальная медицинская библиотека США начала разработку системы Unified Medical Language System (UMLS) - многочисленных словарей, содержащих комплексные тезаурусы и онтологии биомедицинских концептов [1]. В систему UMLS интегрировано около 3,7 млн. концептов с 12 млн. отношений между ними [2].

Для извлечения формализованных данных из неструктурированного клинического текста на английском языке внутри UMLS используется MetaMap [3 Aronson]. На первом этапе работы алгоритма проводится лексический и синтаксический анализ текста. На втором этапе генерируется множество всех возможных сочетаний слов и фраз, полученных на первом этапе, после чего оценивается правдоподобие подобных сочетаний. Данный алгоритм имеет чрезвычайно низкую производительность и с трудом может применяться к большим массивам информации.

При обработке текстов историй болезни, одной из задач является определение связей признаков и характеристик заболеваний с их значениями. Так, например, признак «боль» может иметь характеристику «локализация», которая может принимать значения «поясничная область». В данной работе мы предлагаем новый алгоритм выделения подобных связей.

## 2 Метод выделения связей

Для описания множества возможных признаков, характеристик и их значений использовалась онтология, описанная в [4]. По данной онтологии была разработана база медицинской терминологии и наблюдений. На рис. 1. показан фрагмент этой базы.

Стрелками здесь обозначены отношения “включения”, а точками - отношения “принимают значения”. Группа признаков “Боли” включает признак “Боль в спине” с возможными синонимами: боль в позвоночнике и люмбалгия. Признак “Боль в спине” имеет характеристику “Локализация” с возможными значениями: “поясничная область” и “поясничный отдел позвоночника”, а также характеристику - “усиление” с возможными вариантами “при глубоком вдохе” и “в положении сидя”. У последнего значения характеристики

“усиления”, как и у самой характеристики, имеются синонимы. Синонимы необходимо хранить в базе, так как для обозначения одной и той же сущности в текстах могут использовать различные обозначения. Хотя они и могут быть сходны, формально, они являются различными словосочетаниями, между которыми необходимо поставить соответствие.

Алгоритм основан на интегрированном подходе, объединяющем синтаксический и семантический анализ, когда на одном шаге выполняется синтаксический разбор раздела “Жалобы”, а на другом - результаты разбора верифицируется, в том числе корректируется, на основе семантики описания жалоб. Оба подхода выполняются с использованием формализмов базы медицинских терминов и наблюдений. Для синтаксического анализа она служит языком, т.е. набором лексем, которые могут встретиться в разделе «Жалобы», а для верификации семантика отношений групп признаков, признаков, характеристик и их значений позволяет корректно связывать лексемы и группы лексем между собой и ассоциировать их с единственно возможным Признаком, Характеристикой или их значением.

**Этап 1.** Целью первого этапа является создание упорядоченного набора лексем или сложных/составных токенов, которые представлены в базе Наблюдений. Этап содержит несколько шагов.

Шаг 1. Выделение раздела Жалобы в тексте истории болезни. Самым простым случаем является выгрузка раздела Жалобы из некоторой МИС, где он хранится в текстовом формате. В этом случае мы явным образом имеем текст жалоб для дальнейшего анализа. Если истории болезни хранятся в виде текстового документа, то необходимо выделить этот раздел. Раздел обычно расположен в начале текста истории болезни, начинается с нового абзаца, имеет одну из цепочек вида “Жалобы/ Жалобы при поступлении/ На момент осмотра жалобы”. Раздел содержит один или более абзацев и заканчивается с началом нового раздела. Новый раздел определен в структуре историй болезни, и представляет из себя, например, раздел Анамнез заболеваний/Анамнез жизни и др.

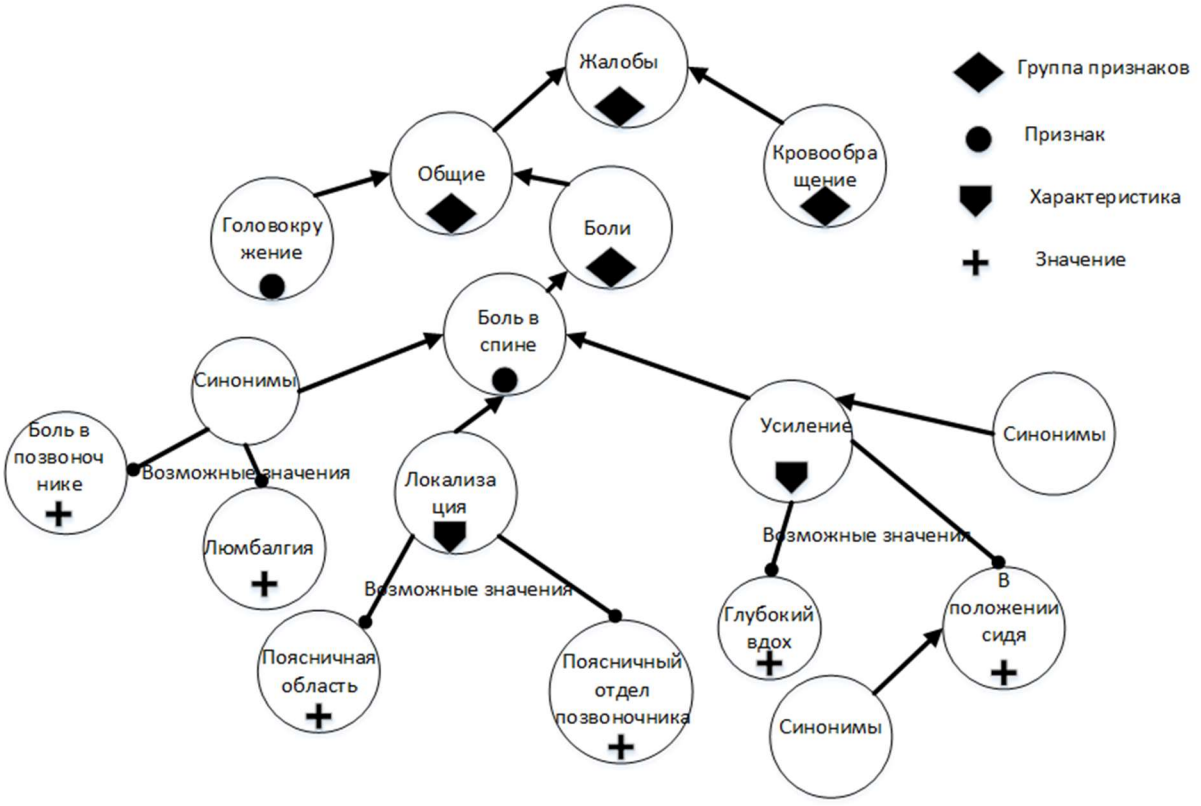


Рис. 1. Фрагмент базы Наблюдений, включая отношения между признаками, характеристиками, синонимами и значениями

В некоторых случаях пациент не предъявляет жалоб в силу возраста, тяжести состояния или их отсутствия. Эти случаи отделяются с помощью поиска цепочек вида “жалоб нет/жалоб не предъявляет/жалоб по возрасту не предъявляет/жалоб не предъявляет в связи с тяжестью” и др.Шаг 2. Выполняется процедура лексического анализа текста выделенного раздела Жалобы. В нашей работе применяется морфологический анализатор русского языка PyMorphy2 [PyMorphy2], использующий словарь OpenCorpora [OpenCorpora]. Результатом является упорядоченный набор отдельных слов в начальной форме, которые необходимо связать с Признаками (*Pi*), характеристиками или их значениями (*Hik*).

Шаг 3. Выполняется процедура установления возможных связей выделенных слов с Признаками, Характеристиками или их значениями. Термины могут состоять из одного или нескольких слов. Выделение терминов выполняется с использованием базы Наблюдений. Среди наименований Признаков, Групп признаков, наименований Характеристик, значений Характеристик и всех возможных синонимов мы выбираем те, которые соответствуют данным раздела Жалобы и имеют наибольшую длину. Алгоритм формирования терминов подробно описан ниже.

Для поиска все наименования Признаков, Групп признаков, наименований и значений Характеристик, а также все возможные синонимы организовываются в префиксное дерево, где узлами дерева служат отдельные слова. Таким образом, задача поиска названия в тексте жалобы сводится к поиску максимальной цепочки слов начиная с текущей позиции, ведущей от корня префиксного дерева к одному из его листьев.

Результатом Этапа 1 является упорядоченный набор слов и словосочетаний, часть из которых является группами признаков, признаками, а часть их характеристиками. Отдельные характеристики могут быть неоднозначно выделены, т.е. представлять из себя пул возможных вариантов, так как в базе Наблюдений разным признакам разрешено иметь одинаковое значение характеристики. В полученном упорядоченном наборе терминов связь между признаками и характеристиками не установлена.

**Этап 2.** Основной задачей Этапа 2 является определение связей терминов, выделенных из текста Жалобы, между собой, а также уточнение связи между терминами и узлами базы Наблюдений. На данном этапе устраняются две основные проблемы. В связи с особенностями русского языка, связи между выделенными в тексте терминами могут быть не очевидны. В связи с этим соединение терминов между собой может пройти некорректно и, например, значение характеристики одной жалобы будет приписано характеристике другой. Второй проблемой является неоднозначность терминов, хранимых в базе Наблюдений. Как уже отмечалось выше, одному и тому же термину может быть приписано несколько значений из базы Наблюдений, так как одно и то же проявление может быть присуще разным жалобам. В связи с этим необходимо выделить минимальный набор узлов из базы Наблюдений, который связан именно с этой жалобой (в идеальном случае узел должен быть единственным). Например, боль в любой части тела может быть выраженной, что отражается наличием соответствующих узлов в базе Наблюдений, но по результатам выполнения Этапа 2 должна быть оставлена единственная связь с выраженностью боли конкретной части тела, описываемой в жалобе.

Удобным является тот факт, что в предложениях на русском языке значения характеристики обычно синтаксически связываются с названием характеристики (если она присутствует в тексте). Неоднозначность вносит тот факт, что синтаксический анализ текстов всё еще не достиг стопроцентной точности. В связи с этим в работу Этапа 2 добавляется несколько шагов, компенсирующих эти недостатки.

Рассмотрим теперь последовательность шагов Этапа 2.

Шаг 1. Сложность проведения синтаксического анализа зависит от длины анализируемого предложения и сложности входящих в него конструкций. В связи с этим мы объединяем многословные термины, хранимые в базе Наблюдений, в одно слово. При этом при помощи синтаксического анализатора из многословного термина выделяется главное слово, и его грамматические характеристики приписываются все термину. Начальной формой такого слова будет вся строка термина целиком. Подобным образом сокращается количество слов в предложении, а также гарантируется, что синтаксический анализ не проведет разбор многословного термина некорректно.

Шаг 2. Преобразованное предложение передается в синтаксический анализ, который

строит дерево зависимостей из слов предложения. Синтаксический анализ учитывает особенности построения предложения, например, связи между терминами при помощи предлогов. Это позволяет не искать связи между Характеристикой и ее Значением в том случае, если синтаксической связи между ними быть не может.

Шаг 3. В связи с тем, что синтаксический анализ может допускать ошибки при построении дерева зависимостей, мы вынуждены корректировать полученный результат. Ошибки происходят, например, в связи с тем, что само предложение может иметь несколько вариантов разбора. Кроме того, предложения в историях болезни строятся по довольно специфичным шаблонам, редко встречающимся в других текстах. Так, в предложении относительно реже встречаются глаголы (а могут и не встречаться совсем). Таким образом, Шаг 3 состоит в корректировке связей между терминами. Для проверки корректности связей используется иерархическая природа построения базы Наблюдений. Так, мы можем считать, что термин, находящийся выше в иерархии Базы, не может быть потомков в дереве зависимостей предложения для термина, находящегося ниже. Связь между терминами в дереве зависимостей может быть построена только если есть прямой путь между терминами в базе Наблюдений.

Перестановки узлов в дереве ведутся по следующим правилам.

Шаг 3.1. В результате синтаксического анализа может получиться ситуация, когда родительская вершина в дереве зависимостей будет иметь более низкий уровень иерархии, чем вершина-потомок. Например, дочерней вершиной будет являться характеристика, тогда как родительской вершиной - ее значение. В подобной ситуации следует поменять вершины местами.

Шаг 3.2. Из-за ошибок синтаксического анализа возможен случай, когда родительская вершина является значением от другой характеристики или параметра. В такой ситуации мы поднимает вершину-потомка на уровень родительской вершины.

Шаг 3.3. Если как родитель, так и потомок, являются параметрами, но при этом параметр-потомок имеет более высокий уровень иерархии, мы меняем вершины местами. Если в базе Наблюдений нет прямого пути между этими двумя параметрами, то параметр-потомок поднимается в дереве зависимостей на один уровень выше.

Шаг 3.4. Сравниваем вершину-параметр с другими вершинами, имеющими того же родителя. Если какая-либо другая вершина находится на более высоком уровне иерархии в базе наблюдений, чем данная, подчиняем текущую вершину найденной.

Шаги применяются до тех пор, пока происходят изменения в дереве.

Применяя данные шаги, мы рассчитываем, что правила путем последовательных перестановок восстановят корректный порядок подчинения извлеченных терминов. Например, если синтаксический анализ перепутал связи и подчинил значения Характеристик, перекрестно перемешав их между собой, то сперва значения поднимутся на уровень выше, а затем свяжутся с характеристиками корректным образом. Так как итоговое дерево должно иметь упорядоченный по иерархии базы Наблюдений вид, то мы поднимаем некорректно подчиненные вершины до тех пор, пока не будет обнаружен уровень, на котором находятся его истинные родительские вершины. В крайнем случае, вершина будет поднята на уровень Жалобы и останется без связей, что может служить сигналом о неполноте базы Наблюдений.

## 3 Результаты экспериментов

Данный алгоритм был реализован на языке Python с применением библиотек PyMorphy [5], использующей словарь OpenCorpora [7] и UDPipe [6]. На вход алгоритму были поданы несколько десятков историй болезни, заполненных врачами-неврологами отделения Неврологии и нейрохирургии МЦ ДВФУ. Перед передачей текстов из них была предварительно удалена персональная информация. Рассмотрим разбор на одном и примеров. Из истории болезни было извлечено следующее описание жалоб: «При поступлении жалобы на выраженную боль в поясничном отделе позвоночника, ограничение движений в пояснично-крестцовом отделе, нарушение ходьбы, нарушение функции тазовых органов».

После выделения терминов и синтаксического анализа предложения было получено следующее дерево зависимостей (следующий уровень иерархии в дереве показан отступами, как F отмечены названия параметров, V — его значение).

```
2 поступлении
    1 при
    3 жалобы [F]
       6 боль [F]
          4 на
          5 выраженную [V]
       8 поясничном отделе позвоночника [V]
          7 в
    9 ,
    10 ограничение движений [F]
       12    пояснично-крестцовом отделе [V]
          11 в
       14 нарушение ходьбы [F]
          13 ,
       16 нарушение функции тазовых органов [F]
          15 ,
```

Как видно, синтаксический анализ некорректно объединил «жалобы» и «в поясничном отделе» (узлы 3 и 8); также узлы 14 и 16 некорректно подчинены узлу 10. Помимо этого, узлы 10, 14 и 16 должны быть также отнесены к жалобам, то есть подчиняться узлу 3. После перестановок и уточнения пути до термина от корня дерева базы наблюдений получается следующее дерево.

```
2 поступлении
    1 при
    3 жалобы ['Жалобы']
       6 боль ['Жалобы| Общие| Боли']
          4 на
          5 выраженную ['Жалобы| Общие| Боли| Боль в спине| Составной признак| Выраженность| Тип возможных значений| Качественные значения| выраженная']
          8 поясничном отделе позвоночника ['Жалобы| Общие| Боли| Боль в спине| Составной признак| Локализация| Тип возможных значений| Качественные значения|поясничный отдел позвоночника']
             7 в
       10 ограничение движений ['Жалобы| Опорно-двигательная система| Снижение подвижности сустава']
          12 пояснично-крестцовом отделе ['Жалобы| Опорно-двигательная система| Снижение подвижности сустава| Составной признак| Локализация| пояснично-крестцовый отдел']
             11 в
       14 нарушение ходьбы ['Жалобы| Нервная система| Нарушение походки']
          13 ,
       16 нарушение функции тазовых органов ['Жалобы| Нервная система| Нарушение функции тазовых органов']
          15 ,
    9 ,
```

Как видно, узлы дерева корректно подчинены друг другу: узлы расположенные на более низком уровне иерархии подчиняются узлам более высокого уровня, сами связи соответствуют синтаксическим связям в предложении.

## 4 Выводы

По отдельности, синтаксический анализ и построение связей между терминами в соответствии с их иерархией в онтологии не дают корректного результата. Синтаксический анализ совершает ошибки соединения соседних терминов (в нашем примере на 16 слов было совершено 3 ошибки синтаксического подчинения). Прямое соединение терминов между собой дает большую неоднозначность, которая не всегда может быть корректно разрешена без использования синтаксического анализа (или как минимум, некоторых предположений о структуре предложения в языке). Таким образом, именно соединение двух подходов позволяет получить быстрое и точное решение.

Однако метод перестановок терминов обладает некоторыми недостатками. Так, при подъеме термина на уровень выше следует сохранять информацию о последовательности слов. Если поднимаемый термин будет размещен в конце списка, он может быть некорректно подчинен родительскому термину, если в данном месте дерева есть несколько терминов с одинаковым префиксом. Кроме того, наш алгоритм пока не строит дерево связей между терминами в онтологии, ограничиваясь лишь построением синтаксического дерева зависисмостей.

## 5 Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-03131.

## Список литературы

1. Selden CR, Humphreys BL. *Unified Medical Language System® (UMLS®).* 1996. https://www.nlm.nih.gov/archive/20040831/pubs/cbm/umlscbm.html

2. Bodenreider O. *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Research, 2004, Vol. 32, Database issue D267-D270.

3. Aronson A.R, Lang F.M. *An overview of MetaMap: historical perspective and recent advances.* Journal of the American Medical Informatics Association. 2010;17(3):229–236. Doi:10.1136/jamia.2009.002733

4. Грибова В. В., Москаленко Ф. М., Шахгельдян К.И., Гмарь Д.В., Гельцер Б.И. *Концепция гетерогенного хранилища биомедицинской информации* // Информационные технологии, Том 25, №2, М.: "Новые технологии", 2019, сс. 97-106

5. Korobov M.: *Morphological Analyzer and Generator for Russian and Ukrainian Languages* // Analysis of Images, Social Networks and Texts, pp 320-332 (2015).

6. Milan Straka and Jana Strakova. Tokenizing, *POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99, Vancouver, Canada, August 3-4, 2017

7. *Сегментация текста в проекте "Открытый корпус"* // "Компьютерная лингвистика и интеллектуальные технологии" Материалы ежегодной Международной конференции «Диалог» 2012, Выпуск 11, Том 1, сс. 51-60