

Московский физико-технический институт (государственный университет)

Факультет биологической и медицинской физики

Кафедра Молекулярной медицины

Диссертация допущена к защите

зав. кафедрой

\_\_\_\_\_ Лазарев В.Н.

«\_\_\_\_\_» \_\_\_\_\_ 2016 г.

**Выпускная квалификационная работа  
на соискание степени  
МАГИСТРА**

**Тема: Автоматический анализ текстов для выявления  
отношений между бактериями микробиоты  
кишечника человека, питанием и заболеваниями**

Направление: 010900 – Прикладные математика и физика

Магистерская программа: 010982 – Физико-химическая биология и биотехнология

Выполнил студент гр. 0112 \_\_\_\_\_ Ярыгин К. С.

Научный руководитель,

к. б. н. \_\_\_\_\_ Лазарев В.Н.

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России

Москва – 2016

## Оглавление

1.	Список сокращений . . . . .	3
2.	Введение . . . . .	4
3.	Литературный обзор . . . . .	6
3.1.	Микробиота кишечника человека . . . . .	6
3.2.	Интеллектуальный анализ текстов . . . . .	9
	<b>Список литературы . . . . .</b>	<b>22</b>

## 1. Список сокращений

ЖКТ — Желудочно-кишечный тракт

ИАТ — Интеллектуальный анализ текстов

ИАБТ — Интеллектуальный анализ биомедицинских текстов

КЖК — Короткоцепочечные жирные кислоты

MEMM, МММЭ — Maximum entropy markov models, Марковские модели максимальной энтропии

SVM, МОВ — Support vector machine, Метод опорных векторов

NLP, ОЕЯ — Natural language processing, Обработка естественных языков

NER, РИС — Named entity recognition, Распознавание именованных сущностей

CRF, УСП — Conditional random fields, Условные случайные поля

CSD — Complementary structures in disjoint literatures, Комплементарные структуры в непересекающихся литературах

GO — Gene ontology, Онтология генов

MeSH — Medical subject headings, Медицинские предметные рубрики

NGS — Next-Generation sequencing, Секвенирование нового поколения

UMLS — Unified medical language system, Унифицированный язык медицинских систем

## 2. Введение

Микробные сообщества, населяющие различные части кишечника человека, могут оказывать влияние на здоровье своего хозяина. В здоровом организме они способствуют пищеварению, расщепляя не переваренные человеком компоненты пищи и синтезируя различные нутриенты и витамины. Негативные эффекты, оказываемые кишечными микробами, включают в себя различные воспаления и инфекции, участие в возникновении и развитии болезней желудочно-кишечного тракта (ЖКТ), а также диабета и ожирения. В последнее десятилетие был достигнут огромный прогресс в методах определения состава кишечной микробиоты и роли бактерий в кишечном метаболизме. Стало понятно, что питание более всего остальных определяет состав микробиоты в краткосрочной и долгосрочной перспективе, что открывает возможности манипулировать состоянием микробиоты меняя диету. Также стала ясна роль межиндивидуальных различий в составе микробиоты, которые вероятно служат причиной различных ответов на прием лекарственных средств или смену диеты. Достижение более точного понимания метаболических ролей различных бактерий в кишечнике, их взаимодействия друг с другом и эффектов, которые это взаимодействие оказывает на человека, очень важно для лечения и предотвращения возникновения болезней с микробиотной компонентой в этиологии.

К настоящему времени накоплено огромное количество научных данных о микробиоте кишечника человека. На данный момент база данных научной литературы Pubmed Central индексирует около 20,000 статей на тему “Микробиота кишечника человека” и с каждым годом это количество увеличивается экспоненциально. Учитывая это обстоятельство, следить за релевантными публикациями становится все сложнее, а поиск нужной информации вручную во всех накопленных источниках и сопоставление ее с результатами экспериментов — процесс, занимающий много времени. Для автоматизации же процесса поиска с помощью компьютера.

недостаточно просто загрузить в него грамматику и словарь. Компьютер, как и человек, должен обладать специализированным знанием чтобы понимать текст. Научная область, занимающаяся обучением компьютера понимать написанный человеком текст называется обработкой естественных языков (ОЕЯ, *англ.*: natural language processing, NLP). Интеллектуальный анализ биомедицинских текстов (ИАБТ, *англ.*: biomedical text mining) — ее подраздел, посвященный обработке научных текстов по

биологии, химии и медицине. Основная цель ИАБТ — вычленить какую-либо биологическую информацию из текста, написанного человеком на естественном языке. В настоящее время существует множество систем ИАБТ, успешно применяемых для нахождения белок-белковых взаимодействий [1], ген-белковых взаимодействий [2], определение мишеней лекарственных препаратов [3] и многое другое.

Целью данной работы является создание алгоритма ИАБТ для сбора информации о связях между бактериями, болезнями и питанием из научной литературы. Для этого были поставлены следующие задачи:

- Разработать алгоритм для нахождения информации о биологических связях между бактериями и болезнями, а также между бактериями и нутриентами в произвольном тексте.
- Применить разработанный алгоритм на текстах научных публикаций на тему “Микробиота кишечника человека”.
- Провести валидацию метода на основе полученных результатов.

### 3. Литературный обзор

#### 3.1. Микробиота кишечника человека

Большинство многоклеточных организмов живут в тесной связи с окружающими микробами, и человек не является исключением. Человеческий организм населен большим количеством разнообразных бактерий, архей, вирусов и одноклеточных эукариотов. Микроорганизмы, сосуществующие вместе со своим хозяином, называются микробиотой. Состав микробиоты и ее влияние на организм хозяина стали объектом пристального изучения в последний десяток лет, с появлением технологий NGS (*англ.*: next-generation sequencing, секвенирование нового поколения). В данный момент бактерии остаются самым изученным компонентом микробиоты, в то время как роль вирусов, архей и эукариот менее ясна. По оценкам, микробиота человека содержит около  $10^{13}$ – $10^{14}$  бактериальных клеток, примерно столько же, сколько содержится в самом человеческом теле [4]. Микробиотой обладают практически все поверхности человеческого тела, которые так или иначе имеют контакт с внешней средой: микробы покрывают кожу, желудочно-кишечный, мочеполовой и дыхательный тракт. Самой заселенной средой является желудочно-кишечный тракт (ЖКТ): 70% всех микробов, населяющих человеческое тело, содержатся в толстом кишечнике [5]. Это вызвано тем, что кишечник обладает большой площадью поверхности ( $\sim 200\text{м}^2$ ) [6], а также содержит множество веществ, которые могут использоваться микробами в качестве питания.

Большинство бактерий, проживающих в кишечнике взрослого человека — это неспорообразующие анаэробы. Наиболее представлены среди них бактерии видов *Bacteroides*, *Bifidobacterium*, *Eubacterium*, *Clostridium*, *Lactobacillus* и *Fusobacterium*, а также различные грамположительные кокки [7]. Кишечные микроорганизмы обладают способностью метаболизировать доступные субстраты, которые могут иметь как внешнее происхождение (пища), так и внутреннее (муцин). Основные субстраты, доступные кишечным бактериям, это крахмал и пищевые волокна, а также другие углеводы, такие как олигосахариды и не поглощаемые человеком сахара. Что касается белков, кишечные бактерии могут использовать ферменты хозяина, муцин, а также отслоившиеся клетки эпителия.

Основываясь на их метаболической активности и продуктах ферментации, ки-

шечные бактерии могут быть поделены на полезных и потенциально патогенных. Положительные эффекты на здоровье со стороны микробиоты включают в себя помощь в пищеварении и усвоении пищи, иммуностимулирующее действие, синтез витаминов, подавление роста потенциальных патогенов и уменьшение холестерина [8–10]. Примерами полезных бактерий, особенно важных для человека, могут служить бактерии видов *Lactobacilli* или *Bifidobacteria*, которые помогают переваривать лактозу людям с ее непереносимостью [11], увеличивают устойчивость к инфекциям [12], а также помогают справляться с воспалительными процессами в кишечнике [13]. Отрицательными же эффектами могут являться синтез микробиотой канцерогенных веществ и токсино [14], диарея [15], запоры [16] или даже болезни печени [17], а также кишечные инфекции. Примерами патогенных бактерий являются *Pseudomonas aeruginosa*, вызывающая инфекционную диарею [18], или *Clostridium difficile*, служащая причиной псевдомембранозного колита [19]

## Микробиота и питание

Микробный состав кишечной микробиоты взрослого человека стабилен и мало изменяется со временем. Влияние определенных пищевых компонентов было исследовано в нескольких работах [20; 21], но, несмотря на то, что многие исследования показали наличие реакции микробиоты на определенный пребиотик, только в нескольких были отмечены временные изменения всего микробного состава [22–24]. Похоже, что состав микробиоты зависит больше от индивидуальных особенностей, чем от диеты [22]. С другой стороны, диета все же оказывает влияние на относительную представленность различных групп бактерий в микробиоте. При смене диеты (равно как и при приеме антибиотиков) изменения в представленности бактерий происходят довольно быстро (в течение нескольких дней) и с такой же скоростью возвращаются к начальным значениям при возврате к привычной диете [22].

Метаболическая активность бактерий в кишечной микробиоте может иметь сильные последствия для хозяина, которые могут быть как положительными, так и пагубными. Метаболизм анаэробных микробных сообществ являет собой сложную систему, где синтрофия (симбиотическое сосуществование) является широко распространенным явлением [25]. Только в малой части случаев можно ассоциировать определенные продукты метаболизма с одной или несколькими бактериями. Обычно метаболи-

ты производятся и поглощаются множеством различных членов сообщества. Состав микробиоты, равно как и доступные для нее субстраты, определяют финальные продукты метаболизма сообщества.

У человека нет ферментов, необходимых для расщепления многих структурных углеводов, содержащихся в употребляемых в пищу растениях, и поэтому они непереваренными доходят до толстого кишечника, где служат субстратом для бактерий. Основные формы углеводов, доступных для кишечных бактерий — это резистентный крахмал и различные некрахмальные поли- и олигосахариды. Расщепление различных структурных углеводов в кишечнике крайне специализированно, например, основным расщепителем крахмала является *Ruminococcus bromii*, а целлюлозы — *Ruminococcus champanellensis*. Продуктом ферментации углеводов кишечными бактериями являются короткоцепочечные жирные кислоты (КЖК), в основном ацетат, пропионат и бутират. Бутират является основным источником для клеток эпителия кишечника, пропионат транспортируется к печени, где участвует в глюконеогенезе, а ацетат попадает в кровоток и используется в липогенезе [26]. Также в результате ферментации неперевариваемых углеводов некоторыми бактериями выделяются вещества, которые могут оказывать противовоспалительное и антиоксидантное действие [27]. В противовес ферментации углеводов, бактериальный метаболизм пищевых жиров и белков в толстом кишечнике может привести к синтезу вредных метаболитов, которые могут быть вредны для человека, например, нитрозаминов [14]. Также микробиота может использовать эндогенные субстраты, такие как человеческие ферменты, клетки эпителия кишечника и муцин (например, *Akkermansia mucinophila* — главный потребитель муцина в кишечнике человека [28])

## Микробиота и болезни

Бактериальный состав микробиоты кишечника может флуктуировать со временем, например, при инфекциях или при приеме антибиотиков. Различные болезни также могут оказывать сильное влияние на представленность бактерий в микробиоте. Сейчас для множества болезней доказана их непосредственная связь с состоянием микробиоты кишечника. Среди них — заболевания с аутоиммунной компонентой, такие как целиакия [29], сахарный диабет 2-го типа [30] и ожирение [31]; воспалительные заболевания кишечника [32] (такие, как болезнь Крона или язвенный колит) и



хроническая диарея [15]. Несмотря на установленную связь между некоторыми болезнями и микробиотой, зачастую роль кишечных бактерий в возникновении и развитии болезней ясна не до конца.

То, насколько важна сбалансированная микробиота, особенно заметно, когда речь идет о последствиях приема антибиотиков. Некоторые исследования показали негативный эффект антибиотиков на кишечную микробиоту [33–35]. Отрицательные последствия приема антибиотиков длятся долгое время даже после прекращения их употребления, что свидетельствует о вредоносных эффектах, вызванных дисбалансом в микробном сообществе. Одно из наиболее известных последствий такого дисбаланса — антибиотик-ассоциированная диарея, вызванная усиленным размножением патогенного вида *Clostridium difficile* [36]. Нормальная микробиота подавляет рост патогенов, конкурируя с ними за субстраты, вырабатывая бактериоцины или разрушающие токсины протеазы. Антибиотики разрушают этот сложный комплекс взаимодействий между микробами, позволяя патогенам резко увеличивать свою численность. Поэтому лечение антибиотик-ассоциированной диареи, вызванной *Clostridium difficile*, с помощью ванкомицина и пробиотических дрожжей является более эффективным по сравнению с лечением только ванкомицином [37].

Несмотря на множество защитных механизмов, представленных в слизистой кишечника, человек все же может подвергнуться заражению различными кишечными патогенами (как вирусами, так и бактериями). Колонизация слизистой кишечника патогенными бактериями влечет за собой воспалительный ответ со стороны организма хозяина с целью сопротивления заражению. Однако воспаление также производит обратный эффект, уменьшая жизнеспособность кишечной микробиоты и позволяя патогену занять освободившиеся ниши. Множество патогенных бактерий используют такой воспалительный подход для максимизации своего инфекционного потенциала: *Citrobacter rodentium* [38], *Salmonella enterica* [39], *Helicobacter troglodytes* [40]. В то же время воспаление также дает возможности для роста оппортунистических патогенов, таких как *Enterobacteriaceae* [38].

### 3.2. Интеллектуальный анализ текстов

Научная литература предоставляет огромное количество информации для исследователей. Эта информация может быть использована для построения гипотез,

для интерпретации результатов экспериментов или для овладения знаниями в той или иной научной области.

Существует множество баз данных, в которых содержится биомедицинская научная литература. Одна из самых крупных из них — PubMed — в данный момент содержит более 25 миллионов ссылок на научную литературу из базы данных MEDLINE, биомедицинских журналов и книг (<http://www.ncbi.nlm.nih.gov/pubmed>). Количество ежегодно публикуемых статей растет экспоненциально [41], вместе с ним растет и количество различных экспериментальных данных, получаемых научными лабораториями во всем мире.

Учитывая экспоненциальный рост количества публикаций, следить за релевантными для ученого публикациями становится все сложнее, не говоря о публикациях из смежных областей. Вычленение нужной информации из баз данных научной литературы и сопоставление ее с результатами экспериментов — процесс, занимающий много времени и требующий аккуратного подбора ключевых слов и составления правильных запросов. Таким образом, ученому зачастую не удастся проанализировать все релевантные статьи на определенную тему.

Интеллектуальный (или машинный) анализ текстов (ИАТ, *англ.* text mining) — это автоматическая обработка текста и его анализ, основанные на методах машинного обучения и обработки естественного языка. В настоящее время ИАТ применяется учеными и исследователями для анализа и обработки научной литературы. С помощью ИАТ становится возможным решить множество различных задач — начиная от выявления белок-белковых взаимодействий [1; 42] до построения базы данных для какой-либо специфической научной области [43].

Из-за того, что текст на естественном языке является весьма неоднородным источником информации, автоматическое вычленение из него необходимой информации — нетривиальный процесс. К настоящему времени ИАТ развился в сложную специализированную дисциплину, в которой обработка текста и машинное обучение сопряжены с анализом данных из различных биологических, химических и медицинских баз данных.

Основная цель ИАТ — вычленить необходимое знание, находящееся в тексте, и представить его в явном и кратком виде; шире — получить новую, ранее неизвестную информацию в результате анализа множества фактов, разбросанных в релевантной

литературе. К сегодняшнему моменту ИАТ, используя факты из баз данных и медикобиологических текстов, успешно применяется для определения молекулярных причин заболеваний [44], ген-белковых [2] и белок-белковых взаимодействий [1; 42], аннотации экспрессии генов [45], определения мишеней лекарственных препаратов [3] и многого другого.

### Общая схема анализа

Можно сказать что основная цель ИАТ — добыть из неструктурированного текста, написанного на естественном языке, не выраженные явно знания и представить их в явной форме. Обычно весь процесс ИАТ биомедицинской литературы можно разбить на четыре части:

1. Информационный поиск (*англ.* Information retrieval) — отбор релевантных задаче текстов для последующего анализа
2. Извлечение информации (*англ.* Information extraction) — системы извлечения информации используются для извлечения из анализируемого текста информацию определенного типа. Особо выделяются два подтипа этой задачи:
  - Распознавание именованных сущностей (*англ.* Named entity recognition) — поиск в тексте слов-имен биологических объектов. Это могут быть названия белков, генов, а так же слов, указывающих на них
  - Распознавание отношений (*англ.* Relation extraction) — поиск в тексте отношений между найденными сущностями.
3. Генерация гипотез (*англ.* Hypothesis generation) — системы генерации гипотез делают какие-либо выводы основываясь на данных, полученных при анализе текста

Различные системы ИАТ в зависимости от решаемой задачи могут включать как все вышеперечисленные пункты, так и сосредотачиваться только на одном из них (например распознавание сущностей).

## Информационный поиск

Информационный поиск — первый шаг ИАТ, цель которого выбрать набор документов, который потом будет подвергнут анализу. В большинстве своем выбор обуславливается научной областью. Например, если целью исследования служит анализ причин определенной болезни, то первым шагом будет получение всех биомедицинских текстов (книги, краткие аннотации или полные тексты статей), касающихся ее. Это может быть сделано с помощью специализированных поисковых систем.

Такие системы как PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>) или PubMed Central (<http://www.ncbi.nlm.nih.gov/pmc/>) используют для поиска систему логических сочетаний запросов (например “ген\_1 AND болезнь\_2”). PubMed извлекает документы из базы данных MEDLINE, но предоставляет только аннотации научных статей, PubMed Central предоставляет доступ к полным текстам статей, но только к тем, к которым издательства предоставили свободный доступ.

Также существуют системы, которые не основываются на системе соответствий запросам. Например GoPubMed [46; 47] (<http://www.gopubmed.com/web/gopubmed/>) классифицирует аннотации статей, используя термины из GO (Gene Ontology) и MeSH (Medical Subject Headings) и организует документы в иерархическую структуру, в которой каждому документу присвоены один или несколько “концептов”. GoWeb [48] (<http://gopubmed.org/web/goweb/>) комбинирует веб-поиск на основе ключевых слов, поиск по онтологиям и ИАТ. Textpresso [49] (<http://www.textpresso.org/>) использует для поиска свою собственную онтологию биологических концептов (клетка, ген и т.д.) и их отношений. iHOP [50] (<http://www.pdg.cnb.uam.es/UniPub/iHOP/>) конвертирует статьи из PubMed в граф из генов и белков, делая поиск статей более интуитивным для человека.

## Распознавание именованных сущностей

Документы, прошедшие отбор могут быть проанализированы на предмет вхождения в них специфических слов. Этот шаг называется распознавание именованных сущностей (РИС, *англ.*: named entity recognition, NER). Именованная сущность — ключевое слово или набор слов, обозначающий определенный предмет или биологический концепт (это может быть название гена, белка, болезни и т.д.) В ходе РИС такие ключевые слова, найденные в тексте, связываются с концептами, которые в свою

очередь, могут упоминаться в тексте под множеством имен. Например, упоминание определенного лекарства должно быть найдено в тексте не только по тривиальному названию химического вещества, но и по синонимам в различных номенклатурах, его торговым названиям и т.д.; поиск упоминаний бактерий должен включать полное название вида, его сокращенное и устаревшие названия.

**Идентификация именованных сущностей** Системы распознавания именованных сущностей можно разделить на три большие группы, в зависимости от подхода, который в них применяется [51].

Первая из них — поиск, основанный на правилах (*англ.*: rule-based approach) [52; 53]. Такие методы используют различные характеристики слов и проверяют, удовлетворяют ли они определенным правилам. Характеристиками слова может служить вхождение в него специальных символов, чисел и букв различного регистра (актуально в случае поиска имен генов и белков), а так же его контекст — части речи и грамматические зависимости. Чаще всего поиск, основанный на правилах, используется в связке с поиском, основанном на словаре.

Вторая группа — поиск, основанный на словаре (*англ.*: dictionary-based approach). Данные системы распознавания именованных сущностей основываются на поиске слов, входящих в заранее составленный словарь терминов. При нахождении сущности она связывается с подходящим термином в какой-либо из баз данных. Обычно словари для поиска составляются на основе различных медикобиологических баз данных (например UMLS, GO, ChemID). Системы, основанные на простом прямом поиске по словарю обладают весьма маленькой чувствительностью из-за высокой вариативности написания имен сущностей [54], поэтому были созданы системы, использующие нечеткий поиск [55; 56] и даже выравнивание имен генов друг на друга [57]. В данный момент подходы, основанные на поиске по словарю, остаются наиболее популярными для аннотации медико-биологической литературы, несмотря на возрастающую популярность машинного обучения в задачах РИС.

Третья группа — системы, использующие машинное обучение (*англ.*: machine learning-based approach). Появление таких методов распознавания именованных сущностей стало возможно благодаря постоянно увеличивающемуся количеству уже размеченных текстов (такие совокупности текстов называются корпусами). Наиболее популярные алгоритмы для РИС в биомедицинской литературе — метод опорных

векторов (МОВ, *англ.*: Support Vector Machine, SVM) [58; 59], марковские модели максимальной энтропии (МММЭ, *англ.*: Maximum Entropy Markov Models, MEMM) [60] и условные случайные поля (УСП; *англ.*: Conditional Random Fields, CRF) [61–63]. Несмотря на то, что алгоритмы машинного обучения имеют хорошие значения точности и чувствительности, их эффективность полностью зависит от качества и состава обучающей выборки.

В настоящее время высокоэффективные системы распознавания именованных сущностей совмещают в себе все перечисленные методы и используют различные алгоритмы для пред- и пост-обработки текстов и результатов.

**Разрешение полисемии** Одна из серьезных проблем распознавания именованных сущностей — разрешение неоднозначности в том случае, если термин имеет несколько смысловых значений (полисемия). Например, “*Casp1*” может означать как ген, так и соответствующей ему белок, “*mice*” может означать в зависимости от контекста или “мышь” или название белка “*MicE*”. “*Streptococcus pneumoniae*” может быть ложно определено как название болезни, а “*Streptococcus pneumoniae* 19A serotype” может быть принято за “*Streptococcus pneumoniae*”. Системы разрешения полисемии обычно обучены на огромном массиве энциклопедических данных (например статьи из Википедии) и анализируют контекст, в котором упомянута сущность [64–66].

## Извлечение отношений

Большинство задач интеллектуального анализа текстов требуют большего, чем просто нахождение в тексте различных сущностей, и включают в себя поиск отношений, связывающих эти сущности. В самом простом случае отношения между биомедицинскими сущностями бинарны и включают в себя только пару сущностей и зависимость между ними. Например, в предложении “*glnAP2* может быть активирован *NifA*” глагол “активирован” определяет тип отношения, слова “*glnAP2*” и “*NifA*” являются его аргументами. Однако биологические отношения могут быть сложны и включать в себя три и более сущности и несколько связей разных типов. Например предложение “стероид эффективно ингибирует фосфорилиацию тирозина у *STAT6*” описывает два биологических отношения, одно из которых определяет существитель-

ное “фосфорилиция”, а другое, сложное отношение, определяемое глаголом “ингибирует” принимает в качестве аргументов сущность “стероид” и первое отношение. Часто отношения между биологическими сущностями также называют “событиями”.

Главная цель извлечения отношений — найти в тексте зависимость между сущностями и определить ее тип. Искомые отношения могут быть как простыми и определенными (как в случае поиска регуляторных отношений между генами), так и сложнокатегоризируемыми, включающими любой тип биомедицинской связи.

В медицине и биологии извлечение отношений имеет множество областей применения. В настоящее время основной задачей извлечения отношений является нахождение белок-белковых взаимодействий, которые играют ключевую роль в структурной и функциональной организации клетки. Другие области, где задействовано извлечение отношений это нахождение взаимодействий между белками и их сайтами связывания, белками и точечными мутациями, генами и болезнями, лекарствами и их мишенями и многое другое [67].

Методы извлечения отношений могут быть разделены на 3 категории: методы, использующие со-встречаемость терминов; методы, основанные на правилах и методы, использующие машинное обучение [68]. В последнее время было разработано множество методов извлечения отношений, которые используют гибридные подходы, сочетающие два или больше метода для достижения большей точности и чувствительности при анализе предложений сложной конструкции, часто встречающихся в биомедицинской литературе.

**Методы на основе со-встречаемости** Анализ со-встречаемости сущностей является простейшим способом нахождения зависимостей. Если сущности часто встречаются вместе, в одном предложении, абзаце или статье, то велика вероятность что они могут быть так или иначе связаны. Например, со-встречаемость терминов “ретинол-связывающий белок 4 (*RBP4*)” и “инсулинорезистентность” в аннотациях статей предполагает функциональную связь между белком и болезнью. Для оценки достоверности используется относительная частота встречаемости двух терминов. Методы, основанные на со-встречаемости легко реализуемы, но, к сожалению, очевидно не способны предоставить информацию о типе отношения, а потому обладают гораздо более низкой точностью по сравнению с другими методами.

**Методы на основе правил** Для методов, основанных на правилах, используются предопределенные шаблоны, находящие отношения и определяющие их принадлежность к тому или иному типу. Такие правила задаются вручную или высчитываются автоматически из размеченных корпусов. Например, в работе Пру [69] для поиска генных взаимодействий были использованы вручную сконструированные паттерны. Пример такого паттерна: “[продукт гена] играет роль [модификатор] [ген]”, описывающий определенный сценарий взаимодействия. Такой паттерн способен найти фразу “белок *Eg1* играет роль репрессора *BicD*”, вычленив событие — репрессия, и два его аргумента: *Eg1* и *BicD*. В PPInterFinder [70], веб-сервисе для нахождения человеческих белок-белковых взаимодействий в аннотациях научных статей искали паттерны в форме “[белок] \* [отношение] \* [белок]”, “[белок] \* [белок] \* [отношение]” и “[отношение] \* [белок] \* [белок]”. Правила могут использовать синтаксическую структуру предложения, морфологию или части речи входящих в него слов. Для автоматического вычленения информации такого рода используются алгоритмы обработки естественных языков (ОЕЯ, *англ.*: natural language processing, NLP). ОЕЯ основана на априорном знании о структуре языка и том, каким образом специфическая биомедицинская информация упоминается в литературе для трансформации предложений, написанных на естественном языке, в структуру, пригодную для автоматической обработки. Анализ таких структур позволяет добиться лучших результатов, чем простой поиск линейных паттернов. В области нахождения зависимостей огромный прогресс был обеспечен использованием таких структур как грамматика составляющих (*англ.*: constituent grammar) и грамматика зависимостей (*англ.*: dependency grammar) [71].

Грамматика составляющих представляет предложение как иерархическую структуру, дерево, листья которого являются единичными словами из этого предложения. Поддеревья, корнями которых служат узлы графа, включающие в себя более одного слова, называются непосредственными составляющими (*англ.*: immediate constituent). В таком виде сложные грамматические единицы складываются из нескольких более простых и не пересекающихся единиц. Сложные грамматические единицы, включающие в себя группу слов, могут принадлежать одной из шести категорий:

1. именная группа — возглавляется существительным
2. группа прилагательного — возглавляется прилагательным



3. наречная группа — возглавляется наречием
4. предложная группа — возглавляется предлогом
5. глагольная группа — возглавляется глаголом
6. предложение

Ниже на рисунке 1 приведен пример дерева грамматики составляющих.

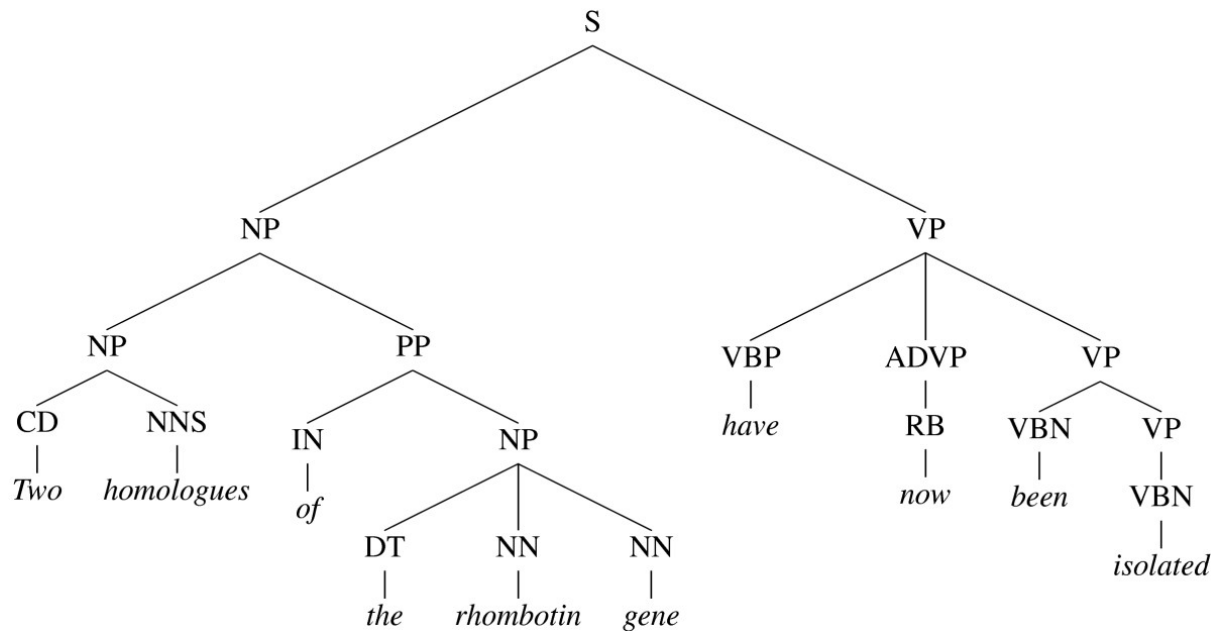


Рис. 1. Дерево грамматики составляющих для предложения “Two homologues of the rhombotin gene have now been isolated” [72]

В отличие от грамматики составляющих (которая более популярна скорее в теоретической лингвистике) грамматика зависимостей гораздо чаще используется в прикладных задачах обработки естественного языка. Грамматика зависимостей представляет предложение в виде графа, узлами которого являются слова из этого предложения, а ребрами — связи между этими вершинами. Связь здесь рассматривается как бинарное асимметричное синтаксическое отношение между двумя словами. На рисунке 2 представлен пример такого графа. Ниже на рисунке 2 приведен пример дерева грамматики составляющих.

Множество современных систем извлечения отношений используют грамматику зависимостей. Например RelEx [73] основана на анализе графа зависимостей с помощью набора простых паттернов, использующих именные группы, части речи слов

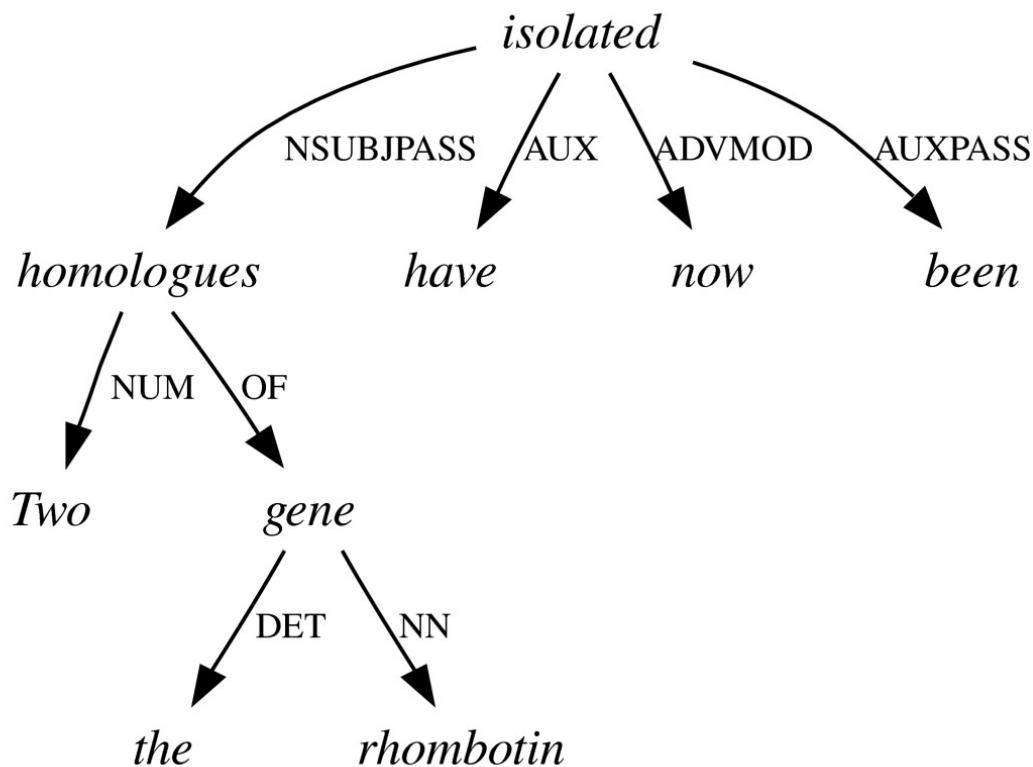


Рис. 2. Граф грамматики зависимостей для предложения “Two homologues of the rhombotin gene have now been isolated” [72]

и поиск отрицаний. RelEx фокусируется на отношениях между генами и белками и использует набор из нескольких паттернов взаимодействия для нахождения таких в тексте. TEES [74; 75] — система извлечения отношений, фокусирующаяся на событиях между белками и генами. TEES использует характеристики графа зависимостей для обнаружения события и метод опорных векторов для его классификации. VExtract [76] основана на анализе кратчайших путей в графе зависимостей между сущностями и ключевыми выражениями.

**Методы на основе машинного обучения** Методы, использующие машинное обучение, воспринимают задачу извлечения отношений как задачу классификации. В таких методах какой-либо классификатор (обычно метод опорных векторов) использует обучающую выборку и основываясь на ней определяет тип отношений. Обычно перед использованием машинного обучения предложения подвергаются анализу: определению частей речи и синтаксическому разбору. Далее получившиеся харак-

теристики используются в качестве признаков объектов (предложений) для обучения классификатора. Также входные данные, подаваемые в классификатор, могут быть представлены в сложной форме, такой как граф зависимостей. Основываясь на формате входных данных, методы, использующие обучение с учителем, могут быть разделены на две группы: методы, основанные на признаках (*англ.*: feature-based methods) и ядерные методы (*англ.*: kernel methods) [68] [77]. Для методов, основанных на признаках, синтаксические и семантические свойства предложения служат подсказками для решения, содержит ли оно какое-либо отношение и, если содержит, то какого оно типа. Часто используемыми свойствами являются входящие в предложения сущности, части речи слов, кратчайший путь между двумя сущностями в графе зависимостей. В [78] для нахождения белок-белковых взаимодействий авторы взяли следующее множество признаков:

- Признаки отдельных слов: слова, входящие в предложение, их леммы и частоты встречаемости, относительное расположение пары белков внутри предложения
- Кратчайший путь в графе зависимостей: слова-вершины графа и ребра-отношения входящие в кратчайший путь
- Граф зависимостей: матрица связности графа.

Каждая группа признаков была агрегирована в вектор и нормализована. В зависимости от набора признаков, данные методы могут быть разделены на поверхностные и глубокие. Первые анализируют информацию не о всем предложении, а только его части, жертвуя полнотой и глубиной анализа, но являются вычислительно-эффективными. Вторые анализируют структуру всего предложения, чаще достигая лучших значений показателей точности и чувствительности, но требуют больших вычислительных ресурсов. Одним из примеров использования методов, основанных на признаках, может служить система PKDE4J [79]. PKDE4J использует автоматически созданные на основе размеченного корпуса текстов паттерны для анализа графов зависимости. EventMine [80] [81] — сервис ИАТ, использующий машинное обучение для нахождения биологических событий в сочетании с поиском со-представленностей для увеличения точности.

Ядерные методы [82] анализируют предложения как строки, используя ядро для извлечения отношений. Для двух строк  $x$  и  $y$  строковое ядро (функция, использую-

щая в качестве аргумента строки) вычисляет их схожесть на основании количества их общих подстрок — чем больше общих подстрок, тем больше схожесть между  $x$  и  $y$ . Каждая строка может быть представлена как вектор в многомерном пространстве, где каждый элемент вектора обозначает либо присутствие (взвешенное,  $(0 : 1]$ ), либо отсутствие (0) определенной подпоследовательности. В более общем представлении  $x$  и  $y$  могут быть любыми объектами, а функция  $K(x, y)$  вычислять их сходство. Так, существуют ядерные функции на подпоследовательностях [83], на деревьях зависимостей [84], на графах зависимостей [85], на кратчайших путях в графах [86] и на комбинации нескольких объектах из вышеперечисленных [87]. Также есть случаи применения гибридных методов. Например Befree [88] — система для идентификации отношений между медицинскими препаратами, генами и ассоциированными с ними болезнями — комбинирует ядерные методы с синтаксическим разбором.

В обоих вышеупомянутых группах методах для обучения классификатора нужна обучающая выборка. В биомедицинском ИАТ одним из самых обширных и наиболее тщательно размеченных корпусов является корпус GENIA [89] (<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA>), состоящий из 2000 аннотаций научных публикаций, полученных по запросам “human”, “blood cells” и “transcription factors”. Разметка включает в себя части речи, синтаксис, кореференцию (указание разных слов на одну и ту же сущность), биологические концепты и события, клеточную локализацию, ассоциации генов с болезнями и метаболические пути. Другие корпуса размеченных аннотаций научных текстов - BioCreAtIve [90–94] и PennBioIE [95]. BioCreAtIve (сокращение от англ. “Critical Assessment of Information Extraction systems in Biology”), (<http://www.biocreative.org/>) - контексты по извлечению отношений в биомедицинских текстах, нацеленные в основном на идентификацию белок-белковых взаимодействий. BioCreAtIve предоставляет размеченный корпус аннотаций научных публикаций к каждому контексту. Корпус PennBioIE содержит 1100 аннотаций статей, посвященных цитохромам *P-450* и 1157 аннотаций статей, связанных с онкологией. Его разметка содержит разделение на параграфы, предложения, токены, части речи слов, синтаксис и биологические сущности.

Также стоит упомянуть о возможных препятствиях, которые могут возникнуть при анализе текстов на предмет обнаружения отношений и событий. Например, найденное в тексте отношение может не только отражать научный факт, но и быть

гипотезой или подвергаться сомнению [96]. Для отделения фактов от гипотез часто требуется ручное курирование результатов анализа. Для того чтобы обойти эту проблему некоторые исследования прибегали к статистическому анализу отношений, которые были извлечены из различных литературных источников и баз данных, с последующим их сравнением между собой [3; 97; 98].

## Список литературы

1. He M., Wang Y., Li W. PPI finder: a mining tool for human protein-protein interactions // PloS one. 2009. Vol. 4, no. 2. P. e4554.
2. Šarić J., Jensen L. J., Ouzounova R. et al. Extraction of regulatory gene/protein networks from Medline // Bioinformatics. 2006. Vol. 22, no. 6. P. 645–650.
3. Chen E. S., Hripcsak G., Xu H. et al. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study // Journal of the American Medical Informatics Association. 2008. Vol. 15, no. 1. P. 87–98.
4. Sender R., Fuchs S., Milo R. Revised estimates for the number of human and bacteria cells in the body // BioRxiv. 2016. P. 036103.
5. Ley R. E., Peterson D. A., Gordon J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine // Cell. 2006. Vol. 124, no. 4. P. 837–848.
6. Gebbers J., Laissue J. Immunologic structures and functions of the gut. // Schweizer Archiv fur Tierheilkunde. 1988. Vol. 131, no. 5. P. 221–238.
7. Wallace T. C., Guarner F., Madsen K. et al. Human gut microbiota and its relationship to health and disease // Nutrition Reviews. 2011. Vol. 69, no. 7. P. 392–403.
8. Renz H., Brandtzaeg P., Hornef M. The impact of perinatal immune development on mucosal homeostasis and chronic inflammation // Nature Reviews Immunology. 2012. Vol. 12, no. 1. P. 9–23.
9. Stecher B., Hardt W.-D. Mechanisms controlling pathogen colonization of the gut // Current opinion in microbiology. 2011. Vol. 14, no. 1. P. 82–91.
10. Smith K., McCoy K. D., Macpherson A. J. Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota // Seminars in immunology / Elsevier. Vol. 19. 2007. P. 59–69.
11. He T., Priebe M., Zhong Y. et al. Effects of yogurt and bifidobacteria supplementation on the colonic microbiota in lactose-intolerant subjects // Journal of Applied Microbiology. 2008. Vol. 104, no. 2. P. 595–604.
12. Liévin-Le Moal V., Servin A. L. Anti-infective activities of lactobacillus strains in the human intestinal microbiota: from probiotics to gastrointestinal anti-infectious biotherapeutic agents // Clinical microbiology reviews. 2014. Vol. 27, no. 2. P. 167–199.
13. Brigidi P., Vitali B., Swennen E. et al. Effects of probiotic administration upon the

- composition and enzymatic activity of human fecal microbiota in patients with irritable bowel syndrome or functional diarrhea // *Research in Microbiology*. 2001. Vol. 152, no. 8. P. 735–741.
14. Gill C., Rowland I. Diet and cancer: assessing the risk // *British journal of nutrition*. 2002. Vol. 88, no. S1. P. s73–s87.
  15. Swidsinski A., Loening-Baucke V., Verstraelen H. et al. Biostructure of fecal microbiota in healthy subjects and patients with chronic idiopathic diarrhea // *Gastroenterology*. 2008. Vol. 135, no. 2. P. 568–579.
  16. Longstreth G. F., Thompson W. G., Chey W. D. et al. Functional bowel disorders // *Gastroenterology*. 2006. Vol. 130, no. 5. P. 1480–1491.
  17. Norman K., Pirlich M. Gastrointestinal tract in liver disease: which organ is sick? // *Current Opinion in Clinical Nutrition & Metabolic Care*. 2008. Vol. 11, no. 5. P. 613–619.
  18. Kim S. W., Peck K. R., Jung S. I. et al. *Pseudomonas aeruginosa* as a potential cause of antibiotic-associated diarrhea. // *Journal of Korean medical science*. 2001. Vol. 16, no. 6. P. 742.
  19. Kelly C. P., Pothoulakis C., LaMont J. T. *Clostridium difficile* colitis // *New England Journal of Medicine*. 1994. Vol. 330, no. 4. P. 257–262.
  20. Zoetendal E. G., Akkermans A. D., De Vos W. M. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria // *Applied and environmental microbiology*. 1998. Vol. 64, no. 10. P. 3854–3859.
  21. Costello E. K., Lauber C. L., Hamady M. et al. Bacterial community variation in human body habitats across space and time // *Science*. 2009. Vol. 326, no. 5960. P. 1694–1697.
  22. Walker A. W., Ince J., Duncan S. H. et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota // *The ISME journal*. 2011. Vol. 5, no. 2. P. 220–230.
  23. Martínez I., Kim J., Duffy P. R. et al. Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects // *PloS one*. 2010. Vol. 5, no. 11. P. e15046.
  24. Russell W. R., Gratz S. W., Duncan S. H. et al. High-protein, reduced-carbo-

- hydrate weight-loss diets promote metabolite profiles likely to be detrimental to colonic health // *The American journal of clinical nutrition*. 2011. Vol. 93, no. 5. P. 1062–1072.
25. Flint H. J., Duncan S. H., Scott K. P., Louis P. Interactions and competition within the microbial community of the human colon: links between diet and health // *Environmental microbiology*. 2007. Vol. 9, no. 5. P. 1101–1111.
  26. Scott K., Duncan S., Flint H. Dietary fibre and the gut microbiota // *Nutrition Bulletin*. 2008. Vol. 33, no. 3. P. 201–211.
  27. Russell W. R., Scobbie L., Chesson A. et al. Anti-inflammatory implications of the microbial transformation of dietary phenolic compounds // *Nutrition and cancer*. 2008. Vol. 60, no. 5. P. 636–642.
  28. Derrien M., Vaughan E. E., Plugge C. M., de Vos W. M. *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium // *International journal of systematic and evolutionary microbiology*. 2004. Vol. 54, no. 5. P. 1469–1476.
  29. De Palma G., Nadal I., Medina M. et al. Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children // *BMC microbiology*. 2010. Vol. 10, no. 1. P. 1.
  30. Larsen N., Vognesen F. K., van den Berg F. W. et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults // *PloS one*. 2010. Vol. 5, no. 2. P. e9085.
  31. Turnbaugh P. J., Hamady M., Yatsunenko T. et al. A core gut microbiome in obese and lean twins // *nature*. 2009. Vol. 457, no. 7228. P. 480–484.
  32. Tannock G. Molecular analysis of the intestinal microflora in IBD // *Mucosal immunology*. 2008. Vol. 1. P. S15–S18.
  33. De La Cochetière M.-F., Durand T., Lalande V. et al. Effect of antibiotic therapy on human fecal microbiota and the relation to the development of *Clostridium difficile* // *Microbial ecology*. 2008. Vol. 56, no. 3. P. 395–402.
  34. Löfmark S., Jernberg C., Jansson J. K., Edlund C. Clindamycin-induced enrichment and long-term persistence of resistant *Bacteroides* spp. and resistance genes // *Journal of Antimicrobial Chemotherapy*. 2006. Vol. 58, no. 6. P. 1160–1167.
  35. Tanaka S., Kobayashi T., Songjinda P. et al. Influence of antibiotic exposure in the early postnatal period on the development of intestinal microbiota // *FEMS Immunology*



- & Medical Microbiology. 2009. Vol. 56, no. 1. P. 80–87.
36. McFarland L. V. Antibiotic-associated diarrhea: epidemiology, trends and treatment. 2008.
  37. Pham M., Lemberg D. A., Day A. S. Probiotics: sorting the evidence from the myths // Medical Journal of Australia. 2008. Vol. 189, no. 3. P. 182–182.
  38. Lupp C., Robertson M. L., Wickham M. E. et al. Host-mediated inflammation disrupts the intestinal microbiota and promotes the overgrowth of Enterobacteriaceae // Cell host & microbe. 2007. Vol. 2, no. 2. P. 119–129.
  39. Stecher B., Robbiani R., Walker A. W. et al. Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota // PLoS Biol. 2007. Vol. 5, no. 10. P. e244.
  40. Whary M., Danon S., Feng Y. et al. Rapid onset of ulcerative typhlocolitis in B6.129P2-IL10tm1Cgn (IL-10-/-) mice infected with Helicobacter trogontum is associated with decreased colonization by altered Schaedler's flora // Infection and immunity. 2006. Vol. 74, no. 12. P. 6615–6623.
  41. Hunter L., Cohen K. B. Biomedical language processing: what's beyond PubMed? // Molecular cell. 2006. Vol. 21, no. 5. P. 589–594.
  42. Zhou D., He Y. Extracting interactions between proteins from the literature // Journal of biomedical informatics. 2008. Vol. 41, no. 2. P. 393–407.
  43. Jamieson D. G., Gerner M., Sarafraz F. et al. Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database // Database. 2012. Vol. 2012. P. bas023.
  44. Srinivasan P., Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases // Bioinformatics. 2004. Vol. 20, no. suppl 1. P. i290–i296.
  45. Draghici S., Khatri P., Martins R. P. et al. Global functional profiling of gene expression // Genomics. 2003. Vol. 81, no. 2. P. 98–104.
  46. Delfs R., Doms A., Kozlenkov A., Schroeder M. GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed. // German Conference on Bioinformatics / Citeseer. Vol. 169. 2004. P. 178.
  47. Doms A., Schroeder M. GoPubMed: exploring PubMed with the gene ontology // Nucleic acids research. 2005. Vol. 33, no. suppl 2. P. W783–W786.
  48. Dietze H., Schroeder M. GoWeb: a semantic search engine for the life science web //

- BMC bioinformatics. 2009. Vol. 10, no. 10. P. 1.
49. Müller H.-M., Kenny E. E., Sternberg P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature // PLoS Biol. 2004. Vol. 2, no. 11. P. e309.
  50. Hoffmann R., Valencia A. Implementing the iHOP concept for navigation of biomedical literature // Bioinformatics. 2005. Vol. 21, no. suppl 2. P. ii252–ii258.
  51. Munkhdalai T., Li M., Kim T. et al. Bio named entity recognition based on co-training algorithm // Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on / IEEE. 2012. P. 857–862.
  52. Fukuda K.-i., Tsunoda T., Tamura A. et al. Toward information extraction: identifying protein names from biological papers // Pac symp biocomput / Citeseer. Vol. 707. 1998. P. 707–718.
  53. Hanisch D., Fundel K., Mevissen H.-T. et al. ProMiner: rule-based protein and gene entity recognition // BMC bioinformatics. 2005. Vol. 6, no. Suppl 1. P. S14.
  54. Tsuruoka Y., Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition // Journal of biomedical informatics. 2004. Vol. 37, no. 6. P. 461–470.
  55. Tsuruoka Y., Tsujii J. Boosting precision and recall of dictionary-based protein name recognition // Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13 / Association for Computational Linguistics. 2003. P. 41–48.
  56. Cohen W. W., Sarawagi S. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining / ACM. 2004. P. 89–98.
  57. Krauthammer M., Rzhetsky A., Morozov P., Friedman C. Using BLAST for identifying gene and protein names in journal articles // Gene. 2000. Vol. 259, no. 1. P. 245–252.
  58. Ju Z., Wang J., Zhu F. Named entity recognition from biomedical text using SVM // Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on / IEEE. 2011. P. 1–4.
  59. Mitsumori T., Fation S., Murata M. et al. Gene/protein name recognition based on

- support vector machine using dictionary as features // BMC bioinformatics. 2005. Vol. 6, no. 1. P. 1.
60. Finkel J., Dingare S., Nguyen H. et al. Exploiting context for biomedical entity recognition: from syntax to the web // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications / Association for Computational Linguistics. 2004. P. 88–91.
  61. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications / Association for Computational Linguistics. 2004. P. 104–107.
  62. Yang L., Zhou Y. Two-phase biomedical named entity recognition based on semi-CRFs // Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on / IEEE. 2010. P. 1061–1065.
  63. Chan S.-K., Lam W. Efficient Methods for Biomedical Named Entity Recognition // 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering / IEEE. 2007. P. 729–735.
  64. Dredze M., McNamee P., Rao D. et al. Entity disambiguation for knowledge base population // Proceedings of the 23rd International Conference on Computational Linguistics / Association for Computational Linguistics. 2010. P. 277–285.
  65. Bunescu R. C., Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation. // EACL. Vol. 6. 2006. P. 9–16.
  66. Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. // EMNLP-CoNLL. Vol. 7. 2007. P. 708–716.
  67. Aggarwal C. C., Zhai C. Mining text data. Springer Science & Business Media, 2012.
  68. Zhou D., Zhong D., He Y. Biomedical relation extraction: from binary to complex // Computational and mathematical methods in medicine. 2014. Vol. 2014.
  69. Proux D., Rechenmann F., Julliard L. et al. A pragmatic information extraction strategy for gathering data on genetic interactions. // ISMB. Vol. 8. 2000. P. 279–285.
  70. Raja K., Subramani S., Natarajan J. PPInterFinder—a mining tool for extracting causal relations on human proteins from literature // Database. 2013. Vol. 2013. P. bas052.
  71. Rastier F., Cavazza M., Abeillé A. Semantics for descriptions: From linguistics to

- computer science. No. 138. Stanford Univ Center for the Study, 2002.
72. Clegg A. B., Shepherd A. J. Benchmarking natural-language parsers for biological applications using dependency graphs // BMC bioinformatics. 2007. Vol. 8, no. 1. P. 1.
  73. Fundel K., Küffner R., Zimmer R. RelEx—Relation extraction using dependency parse trees // Bioinformatics. 2007. Vol. 23, no. 3. P. 365–371.
  74. Björne J., Heimonen J., Ginter F. et al. Extracting complex biological events with rich graph-based feature sets // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task / Association for Computational Linguistics. 2009. P. 10–18.
  75. Björne J., Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task // Proceedings of the BioNLP Shared Task 2013 Workshop. 2013. P. 16–25.
  76. Kilicoglu H., Bergler S. Syntactic dependency based heuristics for biological event extraction // Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task / Association for Computational Linguistics. 2009. P. 119–127.
  77. Bach N., Badaskar S. A review of relation extraction // Literature review for Language and Statistics II. 2007.
  78. Miwa M., Sætre R., Miyao Y., Tsujii J. A rich feature vector for protein-protein interaction extraction from multiple corpora // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 / Association for Computational Linguistics. 2009. P. 121–130.
  79. Song M., Kim W. C., Lee D. et al. PKDE4J: Entity and relation extraction for public knowledge discovery // Journal of biomedical informatics. 2015. Vol. 57. P. 320–332.
  80. Miwa M., Sætre R., Kim J.-D., Tsujii J. Event extraction with complex event classification using rich features // Journal of bioinformatics and computational biology. 2010. Vol. 8, no. 01. P. 131–146.
  81. Miwa M., Thompson P., Ananiadou S. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution // Bioinformatics. 2012. Vol. 28, no. 13. P. 1759–1765.
  82. Lodhi H., Saunders C., Shawe-Taylor J. et al. Text classification using string kernels //

- Journal of Machine Learning Research. 2002. Vol. 2, no. Feb. P. 419–444.
83. Mooney R. J., Bunescu R. C. Subsequence kernels for relation extraction // Advances in neural information processing systems. 2005. P. 171–178.
  84. Moschitti A. Making Tree Kernels Practical for Natural Language Learning. // EACL. Vol. 113. 2006. P. 24.
  85. Airola A., Pyysalo S., Björne J. et al. A graph kernel for protein-protein interaction extraction // Proceedings of the workshop on current trends in biomedical natural language processing / Association for Computational Linguistics. 2008. P. 1–9.
  86. Bunescu R. C., Mooney R. J. A shortest path dependency kernel for relation extraction // Proceedings of the conference on human language technology and empirical methods in natural language processing / Association for Computational Linguistics. 2005. P. 724–731.
  87. Miwa M., Sætre R., Miyao Y. et al. Combining multiple layers of syntactic information for protein-protein interaction extraction // Proceedings of the third international symposium on semantic mining in biomedicine. 2008. P. 101–108.
  88. Bravo À., Piñero J., Queralt-Rosinach N. et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research // BMC bioinformatics. 2015. Vol. 16, no. 1. P. 1.
  89. Kim J.-D., Ohta T., Tateisi Y., Tsujii J. GENIA corpus—a semantically annotated corpus for bio-textmining // Bioinformatics. 2003. Vol. 19, no. suppl 1. P. i180–i182.
  90. Hirschman L., Yeh A., Blaschke C., Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology // BMC bioinformatics. 2005. Vol. 6, no. 1. P. 1.
  91. Krallinger M., Morgan A., Smith L. et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge // Genome biology. 2008. Vol. 9, no. 2. P. 1.
  92. Leitner F., Mardis S. A., Krallinger M. et al. An overview of BioCreative II. 5 // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010. Vol. 7, no. 3. P. 385–399.
  93. Arighi C. N., Lu Z., Krallinger M. et al. Overview of the BioCreative III workshop // BMC bioinformatics. 2011. Vol. 12, no. 8. P. 1.
  94. Arighi C. N., Carterette B., Cohen K. B. et al. An overview of the BioCreative 2012

- Workshop Track III: interactive text mining task // Database. 2013. Vol. 2013. P. bas056.
95. Simpson M. S., Demner-Fushman D. Biomedical text mining: A survey of recent progress // Mining text data. Springer, 2012. P. 465–517.
  96. Névél A., Lu Z. Automatic integration of drug indications from multiple health resources // Proceedings of the 1st ACM international health informatics symposium / ACM. 2010. P. 666–673.
  97. Srinivasan P., Rindflesch T. Exploring text mining from MEDLINE. // Proceedings of the AMIA Symposium / American Medical Informatics Association. 2002. P. 722.
  98. Wang X., Hripcsak G., Markatou M., Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study // Journal of the American Medical Informatics Association. 2009. Vol. 16, no. 3. P. 328–337.