Лаборатория информационных систем
Институт прикладной математики и компьютерных наук
Тульский государственный университет

Богатырев М. Ю.,
Самодуров К. В.

ПРИМЕНЕНИЕ МНОГОМЕРНЫХ ФОРМАЛЬНЫХ КОНТЕКСТОВ В АНАЛИЗЕ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

ММРО-19
26-29 ноября, Москва

План

- Анализ формальных понятий
- Формальные контексты
- Концептуальные графы
- BioNLP:Мультимодальная кластеризация

Анализ формальных понятий

[Wille, 1981]

1. Формальный контекст:

$K = (G, M, R)$

$R \subseteq G \times M$

R – отношение

G – множество объектов

M – множество атрибутов

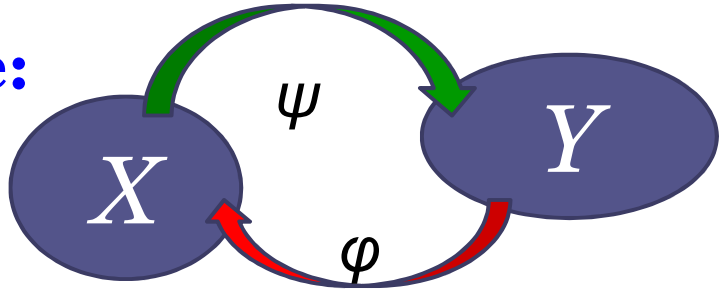
Пример контекста

	Membrane	Nucleus	Replication	Recombination
DNA				X
Virus				X
Prokaryotes	X		X	
Eukaryotes	X	X	X	
Bacterium	X		X	

АФП как метод кластеризации

2. Формальное понятие:

$(X, Y) \quad X \subseteq G,\ Y \subseteq M$



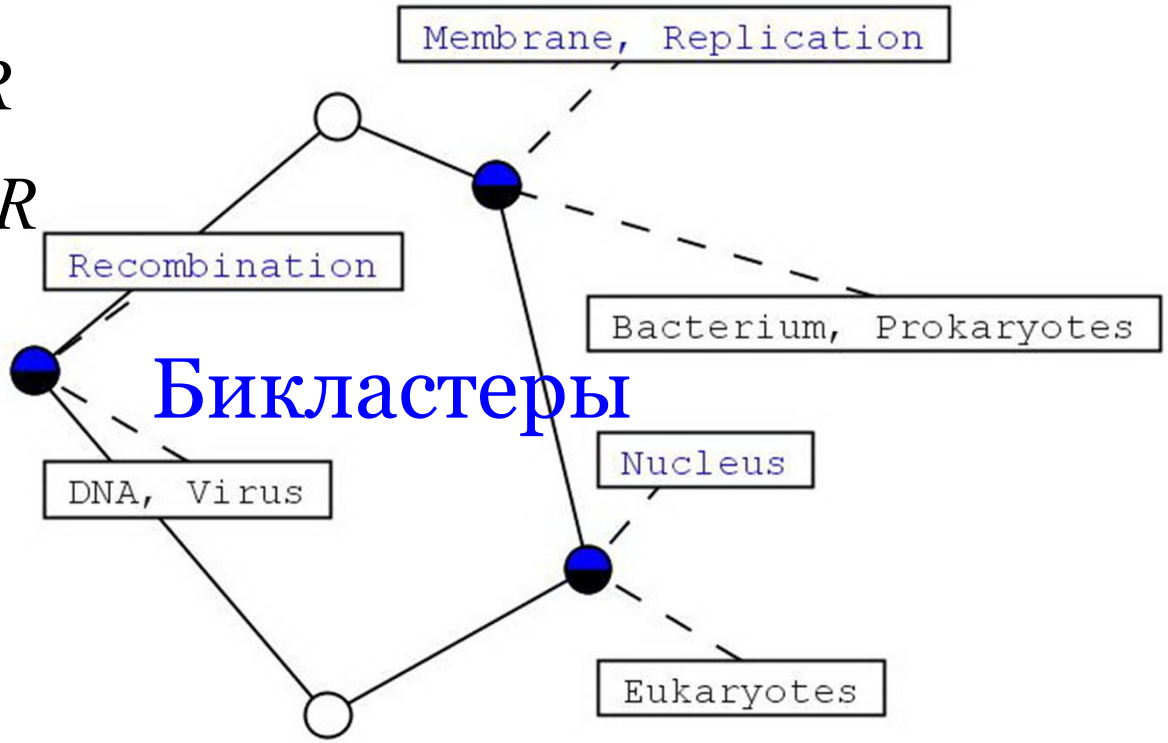
Отображение Галуа

$X \xrightarrow{\psi} Y \qquad Y \xrightarrow{\varphi} X$

$\forall x \in O < x, \psi(x) > \in R$

$\forall y \in A < \phi(y), y > \in R$

3. Решётка понятий:



«Трикластеры» и «трипонятия» в АФП

$\mathrm{K} = (G, M, B, R)$ Тернарный формальный контекст

$R \subseteq G \times M \times B$ R - отношение

$T = (X, Y, Z) = (g^{\square}, m^{\square}, b^{\square})$

$X \subseteq G, Y \subseteq M, Z \subseteq B$ «Трикластер»

$$\rho(X, Y, Z) = \frac{|R \cap X \times Y \times Z|}{|X||Y||Z|}$$ Плотность «трикластеров»

If $\rho(X, Y, Z) = 1$ когда «трикластер" - это «трипонятие»

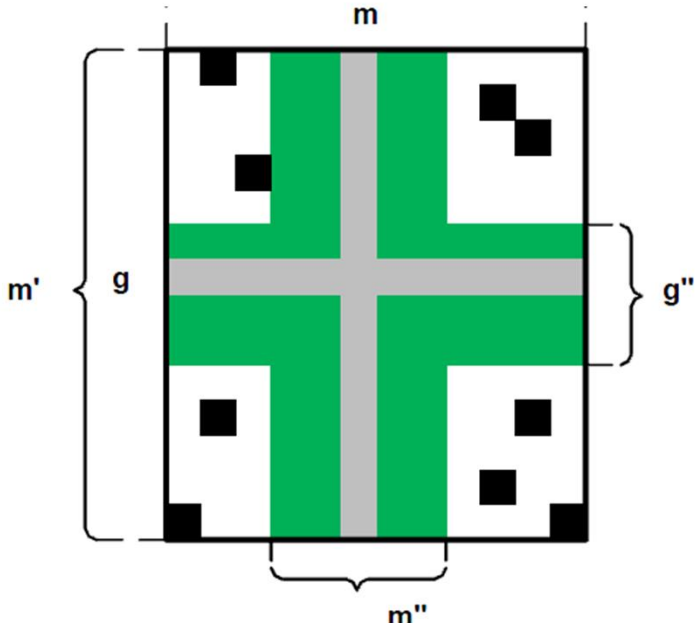
Операторы в OAC-алгоритме трикластеризации

$m' = \{ (g, b) \mid (g, m, b) \in Y \}$	$m'' = \{ \tilde{m} \mid (g, b) \in m' \quad \text{and} \quad (g, \tilde{m}, b) \in Y \}$
$g' = \{ (m, b) \mid (g, m, b) \in Y \}$	$g'' = \{ \tilde{g} \mid (m, b) \in g' \quad \text{and} \quad (\tilde{g}, m, b) \in Y \}$
$b' = \{ (g, m) \mid (g, m, b) \in Y \}$	$b'' = \{ \tilde{b} \mid (g, m) \in b' \quad \text{and} \quad (g, m, \tilde{b}) \in Y \}$

$$g^{\square} = \{ g_i \mid (g_i, b_i) \in m' \text{ or } (g_i, m_i) \in b' \}$$

$$m^{\square} = \{ m_i \mid (m_i, b_i) \in g' \text{ or } (g_i, m_i) \in b' \}$$

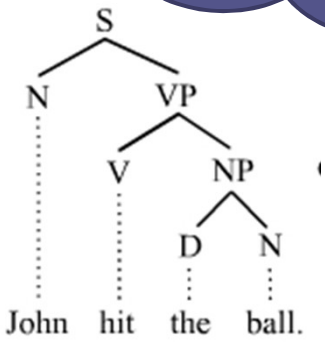
$$b^{\square} = \{ b_i \mid (g_i, b_i) \in m' \text{ or } (m_i, b_i) \in g' \}$$

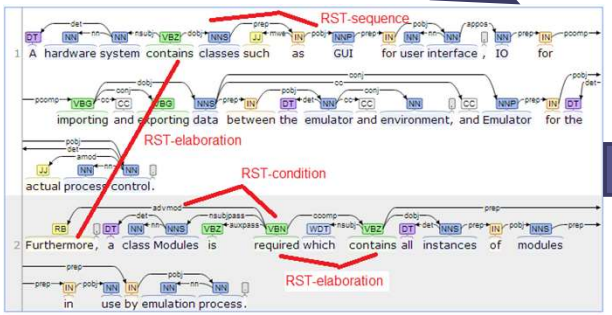


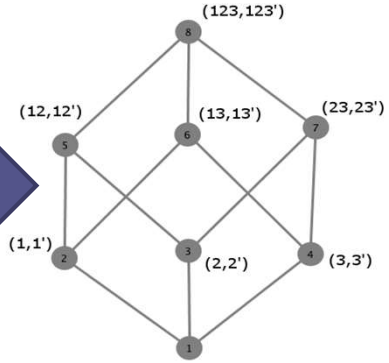
Методы АФП в анализе текстов

Концептуальное шкалирование многозначных контекстов

Теория узорных структур и их проекций







Синтаксические Деревья разбора

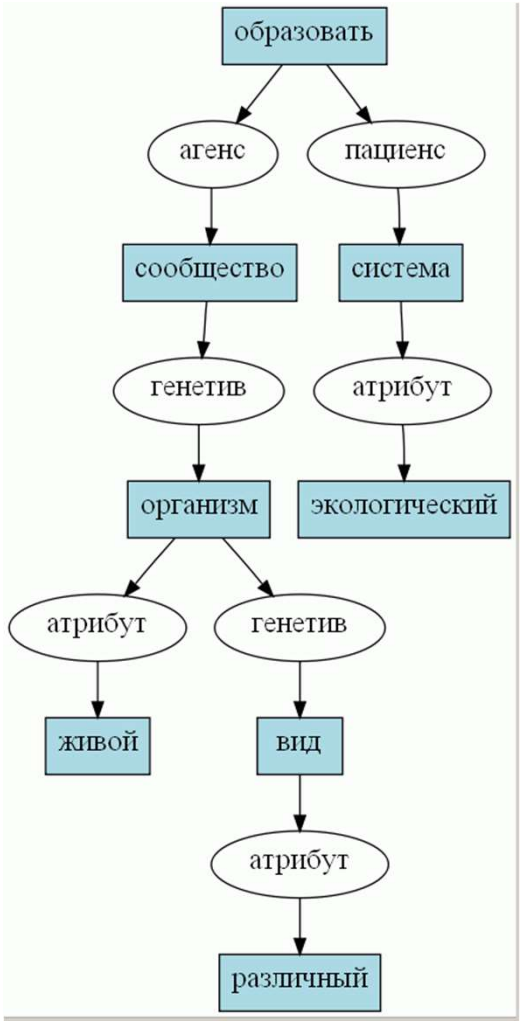
Чащи разбора

Узорная решётка

Ganter, B.; Stumme, G.; Wille, R. Formal Concept Analysis: Foundations and Applications, LNAI, No. 3626, Springer-Verlag. Berlin. 2003.
B. Ganter and S.O. Kuznetsov, Pattern Structures and Their Projections. LNAI, Vol. 2120, pp. 129-142, Springer-Verlag. 2001.

Концептуальные графы [J. Sowa, 1976]

«Сообщества живых организмов различных видов образуют экологическую систему»



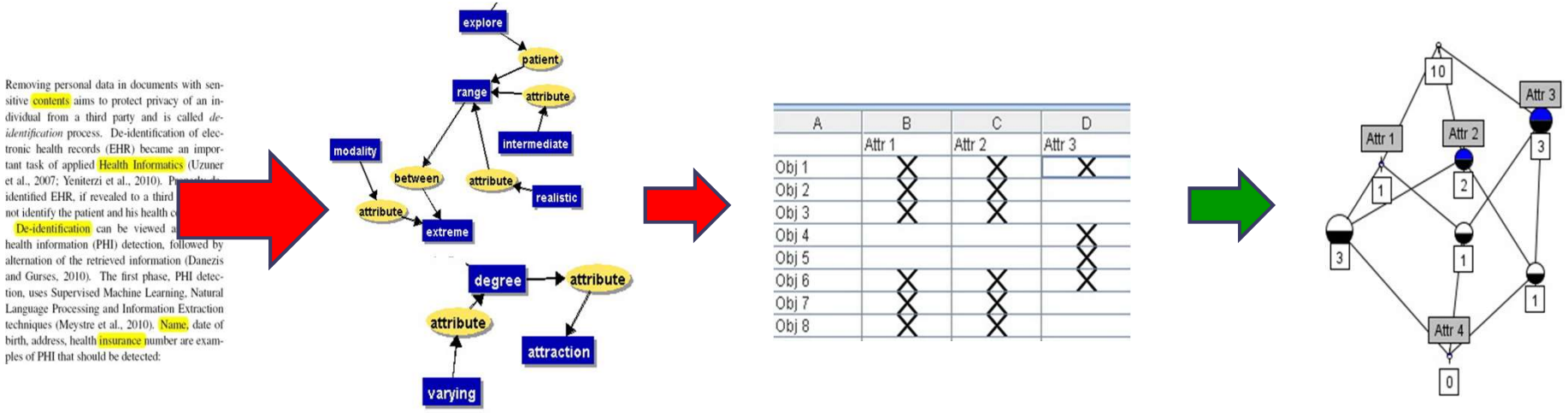
Применение исчисления предикатов:

$$(\exists x : Образовать)(\exists y : Система)(\exists z : Экологический)(\exists w : Сообщество)$$
$$(\exists q : Организм)(\exists g : Живой)(\exists i : Вид)(\exists j : Различный)(Пациенс(x, y) \wedge$$
$$Пациенс(x, z) \wedge Атрибут(y, z) \wedge Генетив(w, q) \wedge Атрибут(g, q) \wedge$$
$$Генетив(i, q) \wedge Атрибут(i, j))$$

Формат xml:

```
- <concepts>
  - <concept number="1">
      <word_norm>сообщество</word_norm>
    </concept>
  - <concept number="2">
      <word_norm>живой</word_norm>
    </concept>
  - <concept number="3">
      <word_norm>организм</word_norm>
    </concept>
  - <concept number="4">
      <word_norm>различный</word_norm>
```

Метод концептуального моделирования: «от слов к понятиям»



Текст ⇒ *концептуальные графы* ⇒ *формальный контекст*

⇓

решётка понятий

Bogatyrev M. Fact Extraction from Natural Language Texts with Conceptual Modeling // Communications in Computer and Information Science, Vol. 706, Springer-Verlag, 2017. — Pp. 89–102.

Гипотеза «семантической выразительности»

Если для моделирования смысла текста
применяется концептуальная модель,
в которой имеет место понятие размерности,
то чем выше размерность такой модели,
тем глубже она позволяет моделировать
смысл текста.

BioNLP

Литература и ресурсы

http://bionlp.org/

Computational Biology
http://www.ploscompbiol.org/

Swanson D.R., Smalheiser N.R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 1997. V. 91. P. 183–203.

Журналы

Journal of Biomedical Semantics
http://www.jbiomedsem.com/
Journal of Biomedical Informatics
http://www.j-biomed-inform.com/home

Основной ресурс

Обзоры

Biomedical Text Mining: a Survey of Recent Progress. In: Mining Text Data. Springer, 2012
Biomedical Text Mining and Its Applications (2009)
http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000597#pcbi-1000597-g002

Конференции

Workshop on Biomedical Natural Language Processing (2019)

Примеры систем BioNLP

1. PIE - PIE (Protein Interaction information Extraction) is a configurable Web service to extract PPI-relevant articles from MEDLINE.
2. KLEIO - an advanced information retrieval system providing knowledge enriched searching for biomedicine.
3. FACTA+ - a MEDLINE search engine for finding associations between biomedical concepts. The FACTA+ Visualizer helps intuitive understanding of FACTA+ search results through graphical visualization of the results.[1]
4. U-Compare - U-Compare is an integrated text mining/natural language processing system based on the UIMA Framework, with an emphasis on components for biomedical text mining.[2]
5. TerMine - a term management system that identifies key terms in biomedical and other text types.
6. PLAN2L — Extraction of gene regulation relations, protein-protein interactions, mutations, ranked associations and cellular and developmental process associations for genes and proteins of the plant Arabidopsis from abstracts and full text articles.
7. MEDIE - an intelligent search engine to retrieve biomedical correlations from MEDLINE, based on indexing by Natural Language Processing and Text Mining techniques [3]
8. AcroMine - an acronym dictionary which can be used to find distinct expanded forms of acronyms from MEDLINE.[4]
9. AcroMine Disambiguator - Disambiguates abbreviations in biomedical text with their correct full forms.[5]
10. GENIA tagger - Analyses biomedical text and outputs base forms, part-of-speech tags, chunk tags, and named entity tags
11. NEMine - Recognises gene/protein names in text
12. Yeast MetaboliNER - Recognizes yeast metabolite names in text.
13. Smart Dictionary Lookup - machine learning-based gene/protein name lookup.
14. TPX - A concept-assisted search and navigation tool for biomedical literature analyses - runs on PubMed/PMC and can be configured, on request, to run on local literature repositories too.[6]

15. Chilibot — A tool for finding relationships between genes or gene products.
16. EBIMed - EBIMed is a web application that combines Information Retrieval and Extraction from Medline.[7]
17. FABLE — A gene-centric text-mining search engine for MEDLINE
18. GOAnnotator, an online tool that uses Semantic similarity for verification of electronic protein annotations using GO terms automatically extracted from literature.
19. GoPubMed — retrieves PubMed abstracts for your search query, then detects ontology terms from the Gene Ontology and Medical Subject Headings in the abstracts and allows the user to browse the search results by exploring the ontologies and displaying only papers mentioning selected terms, their synonyms or descendants.
20. Anne O'Tate Retrieves sets of PubMed records, using a standard PubMed interface, and analyzes them, arranging content of PubMed record fields (MeSH, author, journal, words from title and abtsracts, and others) in order of frequency.
21. Information Hyperlinked Over Proteins (iHOP):[8] "A network of concurring genes and proteins extends through the scientific literature touching on phenotypes, pathologies and gene function. iHOP provides this network as a natural way of accessing millions of PubMed abstracts. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource, bringing all advantages of the internet to scientific literature research."
22. LitInspector — Gene and signal transduction pathway data mining in PubMed abstracts.
23. NextBio- Life sciences search engine with a text mining functionality that utilizesPubMed abstracts (ex: literature search) and clinical trials (example) to return concepts relevant to the query based on a number of heuristics including ontology relationships, journal impact, publication date, and authorship.
24. The Neuroscience Information Framework (NIF) — A neuroscience research hub with a search engine specifically tailored for neuroscience, direct access to over 180 databases, and curated resources. Built as part of the NIH Blueprint for Neuroscience Research.
25. PubAnatomy — An interactive visual search engine that provides new ways to explore relationships among Medline literature, text mining results, anatomical structures, gene expression and other background information.

26. PubGene — Co-occurrence networks display of gene and protein symbols as well asMeSH, GO, PubChem and interaction terms (such as "binds" or "induces") as these appear in MEDLINE records (that is, PubMed titles and abstracts).
27. Whatizit - Whatizit is great at identifying molecular biology terms and linking them to publicly available databases.[9]
28. XTractor — Discovering Newer Scientific Relations Across PubMed Abstracts. A tool to obtain manually annotated,expert curated relationships for Proteins, Diseases,Drugs and Biological Processes as they get published in PubMed.
29. Medical Abstract — Medical Abstract is an aggregator for medical abstract journal fromPubMed Abstracts.
30. MuGeX — MuGeX is a tool for finding disease specific mutation-gene pairs.
31. MedCase — MedCase is an experimental tool of Faculties of Veterinary Medicine and Computer Science in Cluj-Napoca, designed as a homeostatic serving system with natural language support for medical applications.
32. BeCAS — BeCAS is a web application, API and widget for biomedical concept identification, able to annotate free text and PubMed abstracts.
33. @Note — A workbench for Biomedical Text Mining (Including Information Retrieval, Name Entity Recognition and Relation Extraction plugins)

Задачи BioNLP

Извлечение именованных сущностей

Извлечение отношений между сущностями

…

Извлечение фактов

Подходы к решению

Алгоритмические

Технологические

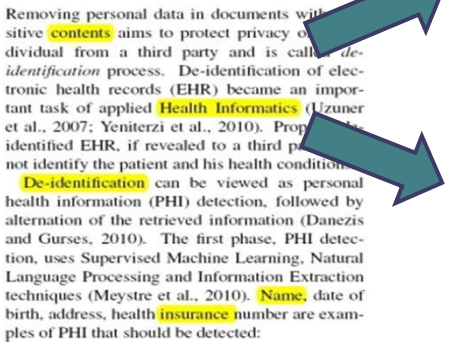
Извлечение именованных сущностей

Название задачи: Biomedical Named Entity Recognition (NER)

Суть задачи: автоматическая диагностика встречаемости биомедицинских терминов в неструктурированных текстах

Вариант: «автоматическое извлечение биомедицинских терминов из неструктурированных текстов».

Вход (тексты)


Removing personal data in documents with
sitive contents aims to protect privacy o
dividual from a third party and is calle
identification process. De-identification of elec-
tronic health records (EHR) became an impor-
tant task of applied Health Informatics (Uzuner
et al., 2007; Yeniterzi et al., 2010). Prop
identified EHR, if revealed to a third p
not identify the patient and his health condition
De-identification can be viewed as personal
health information (PHI) detection, followed by
alternation of the retrieved information (Danezis
and Gurses, 2010). The first phase, PHI detec-
tion, uses Supervised Machine Learning, Natural
Language Processing and Information Extraction
techniques (Meystre et al., 2010). Name, date of
birth, address, health insurance number are exam-
ples of PHI that should be detected:

Выход

Известные термины

Новые термины

Извлечение именованных сущностей

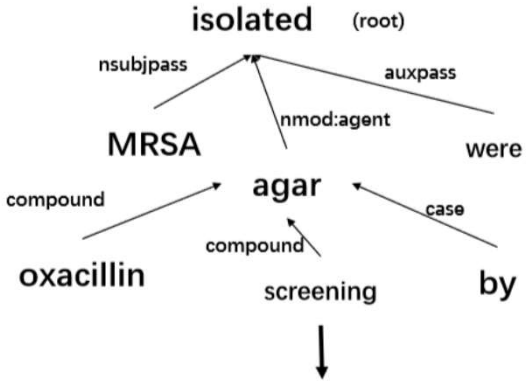
Чаще всего интересуются терминами:

- названия генов,
- названия белков,
- названия лекарств, заболеваний

Извлечение отношений: графовые модели

Графы зависимостей.

MRSA were isolated by oxacillin screening agar.



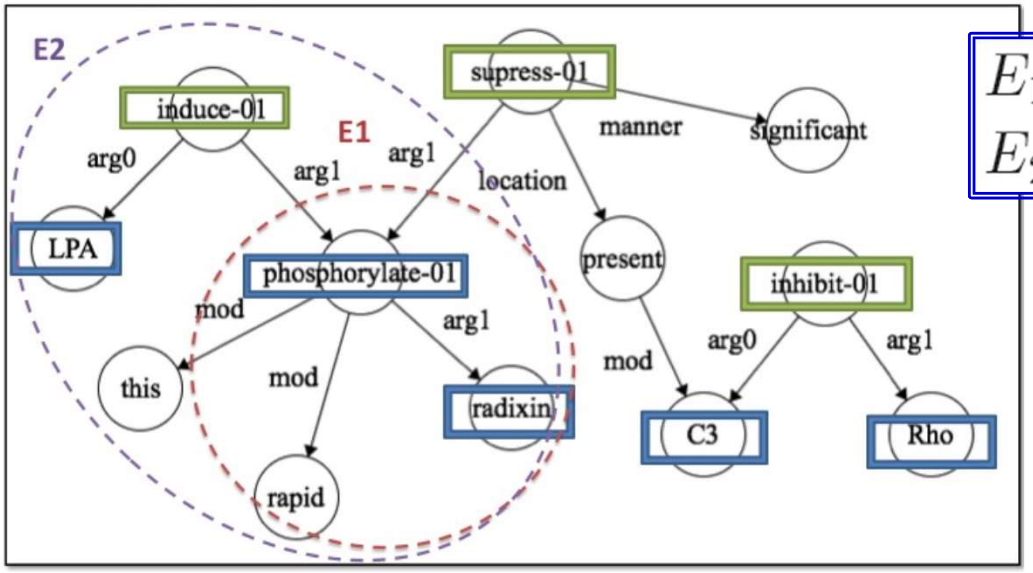
	MRSA	were	isolated	by	oxacillin	screening	agar
MRSA	0	0	1	0	0	0	0
were	0	0	1	0	0	0	0
isolated	1	1	0	0	0	0	1
by	0	0	0	0	0	0	1
oxacillin	0	0	0	0	0	0	1
screening	0	0	0	0	0	0	1
agar	0	0	1	1	1	1	0

Wuti Xiong, Fei Li, Ming Cheng, Hong Yu, Donghong Ji: Bacteria Biotope Relation Extraction via Lexical Chains and Dependency Graphs. Proc. 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China, November 4, 2019 . P.p. 158–167.

Извлечение отношений: графовые модели

Abstract Meaning Representation (AMR)

"This LPA-induced rapid phosphorylation of radixin was significantly suppressed in the presence of C3 toxin, a potent inhibitor of Rho"



$E_1 = phosphorylation\ of\ radixin;$
$E_2 = LPA\ induces\ E1.$

AMR-схемы

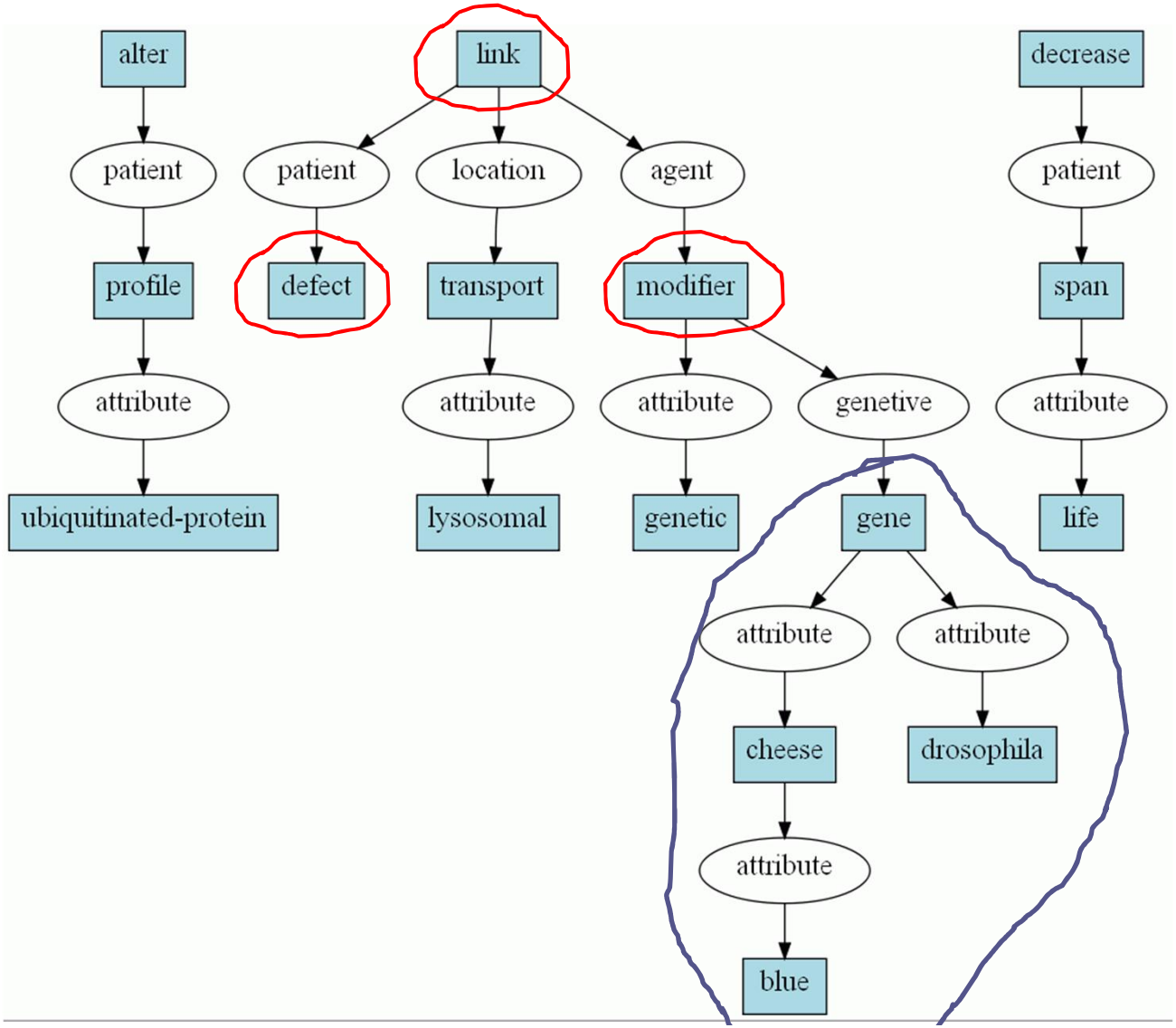
"when"
"how"
....

"who did what to whom"

Sudha Rao, Daniel Marcu, Kevin Knight, Hal Daum´e III. Biomedical Event Extraction using Abstract Meaning Representation. Proc. BioNLP 2017 workshop, Vancouver, Canada. P.p. 126–135, 2017

AMR – схема как подграф концептуального графа

Genetic modifiers of the Drosophila blue cheese gene link defects in lysosomal transport with decreased life span and altered ubiquitinated-protein profiles.



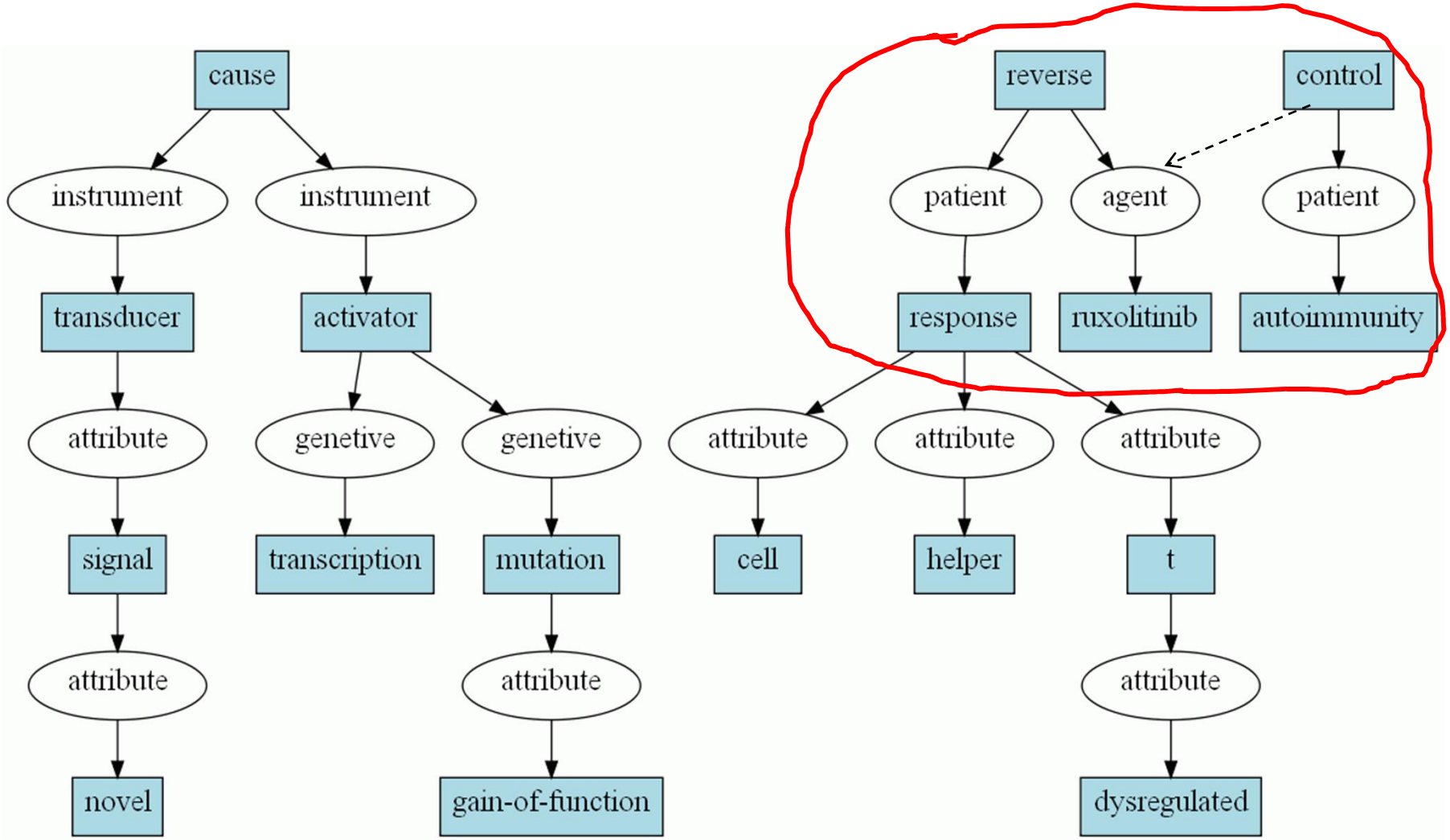
AMR scheme

“*who - did what - to whom*”

“modifier link defect”

Особенности извлечения AMR – схем

“Ruxolitinib reverses dysregulated T helper cell responses and controls autoimmunity caused by a novel signal transducer and activator of transcription of gain-of-function mutation”.



Эксперименты и результаты.

Экспериментальные данные: BioNLP Shared Tasks, https://2019.bionlp-ost.org

Фрагмент разметки:

```
Id   Subject Object  Predicate   Lexical cue
T1   108-117 Var denotes Mutations
T2   121-126 Gene    denotes SHP-2
T3   127-138 Enzyme  denotes phosphatase
T4   150-165 PosReg  denotes hyperactivation

R2 T2   T1   ThemeOf SHP-2,Mutations
R3 T1   T4   CauseOf Mutations,hyperactivation

.....

T4   150-165 PosReg  denotes hyperactivation
 R23     T33 T4   ThemeOf leukemias,hyperactivation
 R24     T34 T4   ThemeOf juvenile myelomonocytic leukemia,hyperactivation
 R3 T1   T4   CauseOf Mutations,hyperactivation
 R4 T5   T4   ThemeOf catalytic activity,hyperactivation
```

Именованные сущности:

- gene
- Enzyme
- PosREg
- .....

Отношения:

- ThemeOf
- CauseOf

Трикластеризация

250 текстов
1787 трикластеров
667 однословных трикластеров

Содержимое кластеров

КЛАСТЕРЫ:{{{cause}, {mutation}, {hyperactivation, ns, jmml, effect, dysplasia,
activation, loss, congenita, phenotype, dementia, ftd, hyperalgesia, osteopenium, osteopenium, nature}},
{{characterize}, {mutation}, {il, presentation, consequence}}, {{induce, promote}, {mutation}, {malignancy,

Сортировка по элементам

‹| mutation → {{{cause}, {mutation}, {hyperactivation, ns, jmml, effect, dysplasia, activation}}, •••••
{{characterize}, {mutation}, {il, presentation, consequence}}, {{induce, promote}, {mutation}, {malignancy}},
{{show}, {mutation, tumor, craniosynostos}, {study, phosphorylation}}, {{play, play},
{mutation, partner, function, regulator, cell, mutation, inhibitor, finding, it, medakon, tissue}, {role, role}}

4-кластеризация

250 текстов
1494 4-кластеров
583 однословных
4-кластера

AMR схема:

“*who - did what - to whom - how*”

Содержимое кластеров

```
{{mutation}, {phosphatase}, {signaling}, {interleukin-3}}, {{mutation}, {cause}, {hyperactivation }, {activity}},
{{mutation}, {induce}, {malignancies}, {hematopoietic}}, {{mutation}, {cause}, {disease}, {myeloproliferative}},
{{mutation}, {increase}, {interaction}, {SHP-2}}, {{mutation}, {increase}, {hyperactivation}, {Erk}},
{{mutation}, {lead}, {hyperactivation}, {pathway}}, {{mutation}, {enhance}, {capability}, {increase}},
{{mutation, SHP-2 E76K}, {enhance}, {activation}, {pathway}}, {{mutation}, {lead, resulted}, {life}, {adult}},
{{mutation}, {lead}, {death}, {neuronal}}, {{mutation}, {lead}, {degeneration}, {CNS}},
{{mutation, mutation}, {defects, defects}, {proteins, proteins}, {motor, cytoskeletal, motor, cytoskeletal}},
{{mutation, mutation}, {behave, behave}, {modifiers, modifiers}, {phenotype, phenotype}},
```

Сортировка по элементам

```
<| {blue cheese gene} → {{{blue cheese gene}, {link}, {defects}, {transport}},
   {{blue cheese gene}, {link}, {span}, {life}}, {{blue cheese gene}, {link}, {profiles},
```

Особенности реализации OAC-алгоритма

3-кластеризация:

```
p3 = Position[data, {i3, j3, _}]; (*находим в списке data позиции *)
p2 = Position[data, {i2, _, j2}]; (*элементов, удовлетворяющих шаблонам*)
p1 = Position[data, {_, i1, j1}];
```

• • • •

```
Do[AppendTo[s3, data[[p3[[n]]]]], {n, Length[p3]}];
(*s3 - список элементв, извлеченных из известных позиций*)
s3 = Flatten[s3, 1];
```

4-кластеризация:

```
p4 = Position[data, {i4, j4, k4, _}];
p3 = Position[data, {i3, j3, _, k3}];
p2 = Position[data, {i2, _, j2, k2}];
p1 = Position[data, {_, i1, j1, k1}];
```

Далее

5-кластеризация?

Анализ ошибок

Error types:

- Ошибки построения концептуальных гра
- Ошибки обработки графов для построения формальных контекстов

«Странные кластеры»:

{{be},{country},{glycoprotein}},{{allow},{country},{delineation}}

play->{{{changes},{play},{leukemigenesis.},{SHP-2-related}}}

Выводы

1. Концептуальный граф – избыточная модель для построения AMR-схем для одного предложения. Однако множество концептуальных графов можно использовать для построения обобщённых AMR-схем для всего текста.
2. Понятие плотности кластера не актуально для «текстовых» кластеров. Важнее фиксировать однословные и неоднословные кластеры
3. Гипотеза семантической выразительности верна, по-крайней мере, для размерностей 3 и 4 многомерных формальных контекстов, построенных на концептуальных графах

Спасибо за внимание!