ПРОБЛЕМА РАСПОЗНАВАНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ В БИОМЕДИЦИНСКИХ ПУБЛИКАЦИЯХ

А. В. Пашук
Ассистент кафедры информатики, магистр технических наук

А. Б. Гуринович
Доцент кафедры вычислительных методов и программирования, кандидат физико-математических наук, доцент

Н.А. Волорова
Заведующая кафедой информатики БГУИР, кандидат технических наук, доцент

А. П. Кузнецов
Прорекор по научной работе, доктор технических наук, профессор

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: pashuk@bsuir.by

Abstract. The number of publications in biomedicine published and indexed annually by PubMed [1] almost doubled over the past 10 years (from 746 thousand to 1354 thousand). This leads to a deterioration in the quality of search and cataloging of scientific publications and it becomes increasingly difficult for scientists to find the necessary information. There is a need to transform unstructured scientific texts into structured formats (XML, JSON). In this task, the quality of recognition of named entities in textual information is of great importance.

Распознавание именованных сущностей (Named-Entity Recognition, сокращенно NER) – это одна из задач извлечения информации из неструктурированного текста, которая заключается в поиске и классификации именованных сущностей, таких как люди, организации, протеины, гены и т.д. Существует целый ряд различных решений данной задачи, большинство из которых позволяет добиться 90% величины F-меры (метрика оценки качества классификатора, объединяющая другие метрики – полноту и точность). Однако, сказанное выше применимо только к классификации ненаучных источников информации. Научная литература и, в частности, биомедицинские статьи и публикации имеют ряд особенностей, которые не позволяют получить хороших результатов с использованием классические алгоритмы распознавания:

– большое количество сокращений, которые могут принимать различные значения, в зависимости от контекста;
– различные варианты написания терминов (например, TNFα, TNF α или TNF-α);
– большое количество синонимов (например, NaCl (соль) имеет больше 300 синонимов [5]);
– наличие двоякости (так называемые омографы), например, CAT (ген) и cat (кошка);
– корректное определение границ термина.

Результат работы классификатора можно представить матрицы неточностей, приведенных в таблице 1.

Зная значения данной матрицы можно посчитать полноту (recall) и точность (precision) поиска – основные метрики, позволяющие оценить качество работы алгоритма поиска.

Таблица 1. Матрица ошибок

		Размеченное значение	
		Правильно	Неправильно
Реальное значение	Правильно	True Positive (TP)	False Negative (FN)
	Неправильно	False Positive (FP)	True Negative (TN)

$$recall = \frac{TP}{TP + FP} \quad (1)$$

$$precision = \frac{TP}{TP + FN} \quad (2)$$

Чтобы улучшить качество работы классификатора именованных сущностей используются различные правила нормализации. Для оценки качества нормализации используются две основных метрики: двоякость (ambiguity) и вариативность (variability) терминов.

Если представить словарь как список терминов $\{t_1, \dots, t_N\}$, где каждый термин связан с идентификатором понятия $c_j \in \{c_1, \dots, c_M\}$. В этом случае двоякость и вариативность терминов можно выразить следующими выражениями [1]:

$$ambiguity = \frac{1}{N}\sum_{i=1}^{N} C(t_i), \quad (3)$$

где N - количество терминов в словаре (онтологии);

$C(t_i)$ - количество понятий в словаре, которые включают в себя слова, по написанию совпадающие с термином t_i.

$$variability = \frac{1}{M}\sum_{j=1}^{M} T(t_j), \quad (4)$$

где M - количество понятий в словаре (онтологии);

$T(T_j)$ - количество уникальных терминов, которые включает понятие c_j.

В таблице 2 приведен пример термина, определенного в нескольких онтологиях (CheBI, DrugBank и др.) и имеющего различные значений в каждой из онтологий.

Нормализацию необходимо использовать только в случае, если она уменьшает одну из метрик (1) или (2), при этом не увеличивая другую. Также показателем качества нормализации служат увеличение точности (precision) и полноты (recall) поиска. Улучшить эти метрики можно, уменьшив количество ошибок первого рода (false positives) или ошибок второго рода (false negatives).

Одним из методов нормализации, позволяющим уменьшить количество ошибок второго рода является использование так называемых генераторов вариантов терминов (Term Variant Generator, сокращенно TVG). Такие генераторы позволяют учитывать не только документы из онтологии, но также различные варианты эти терминов. Например, аббревиатура протеина TNFα может быть записана как TNF α или TNF-α, при этом в онтологии как правило существует один вариант написания (в некоторых случаях онтологии имеют списки синонимов или популярные варианты написания понятия). С использованием TVG будут сгенерированы дополнительные варианты для поиска в онтологии, согласно заложенным правилам.

Таблица 2. Пример термина, имеющего несколько значений

Термин	Онтология	Совпадение в онтологии (ID)	Синонимы термина из онтологии
IMP	ChEBI	Inosine Monophosphate (D007291)	ribosylhypoxanthine monophosphate, inosinic acid, **IMP**, inosinate, sodium, sodium inosinate, inosine monophosphate, acids, inosinic, monophosphate, ribosylhypoxanthine, inosinic acids, monophosphate, inosine, acid, inosinic
	DrugBank	Imipenem (DB01598)	imipemide, imipenem anhydrou, imipenem anhydrous, n-formimidoylthienamycin, imipenem, **IMP**, imipenem and cilastatin for injection, USP, ran-imipenem-cilastatin, imipenem, n-formimidoyl thienamycin, imipenem and cilastatin, imipenemum, imipenem and cilastatin for injection, -USP, primaxin 250, imipenem and cilastatin for injection USP, (5R,6S)-6-((R)-1-Hydroxyethyl)-3-(2-(iminomethylamino) ethylthio)-7-oxo-1-azabicyclo(3.2.0) hept-2-ene-2-carbonsaeure, primaxin IV 500, primaxin 500, primaxin IV 250/250 add-vantage vial, imipenem and cilastatin for injection, usp, imipenem and cilastatin for injection,-usp, (5R,6S)-3-(2-formimidoylamino-ethylsulfanyl)-6-((R)-1-hydroxy-ethyl)-7-oxo-1-aza-bicyclo[3.2.0] hept-2-ene-2-carboxylic acid, imipenem and cilastatin for injection, tienamycin, imipenem and cilastatin for injection usp, imipenem and cilastatin for injection-USP, imipenem and cilastatin for injection-usp, primaxin IV, primaxin-iv, n-formimidoyl thienamycin, (5R,6S)-3-((2-(formimidoylamino) ethyl) thio)-6-((R)-1-hydroxyethyl)-7-oxo-1-azabicyclo(3.2.0) hept-2-ene-2-carboxylic acid
	GeneOntology	obsolete mitochondrial inner membrane peptidase activity (GO:0004244)	**IMP**, obsolete mitochondrial inner membrane peptidase activity, mitochondrial inner membrane peptidase activity
		mitochondrial inner membrane peptidase complex (GO:0042720)	**IMP**, mitochondrial inner membrane peptidase complex
	ChEBI	IMP (CHEBI:17202)	**IMP**, C10H13N4O8P
	Uniprot (Для нескольких организмов)	Inositol monophosphatase (IMPA1_DICDI, IMPP_MESCR)	IMPase, **IMP**, inositol-1(or 4)-monophosphatase, inositol monophosphatase, d-galactose 1-phosphate phosphatase

Рассмотренный в [4] генератор вариантов предусматривает создание вариантов по следующим правилам: преобразование термина во множественный/единичный вид, удаление/добавление знаков пунктуации (таких как дефисы или апострофы). В рамках исследования был разработан собственный алгоритм, расширяющий список правил стандартного генератора для лучшего определения биомедицинских терминов:

– преобразование чисел в начале термина (5-iodotubercidin, 5iodotubercidin, 5 iodotubercidin и т.д.);

– преобразование числе в конце термина (IL-1, IL 1, IL1 и т.д.);

– буквенные индексы в конце термина (penicillin G, penicillin-G и т.д.);

– преобразование греческих символов в их текстовой выражение (TNF α, TNF alpha) и др.

Гистограммы количества оригинальных вариантов (синонимы, варианты написания и т.д.) и количества сгенерированных вариантов (включая оригинальные) изображены на рисунке 1.

Описанные выше правила позволяют классификатору корректно определить и разметить термины в биомедицинской статье или публикации. Также стоит отметить, что из всей массы полученных вариантов в публикациях обычно встречается не больше 30%. Поэтому после генерации вариантов полученный алгоритм проверяет их на текстах статей, отбрасывая варианты, который ни разу не встречаются. Такой подход позволяет уменьшить время и нагрузку на систему, затрачиваемые на разметку текста.

Отдельно стоит рассмотреть вопрос влияния генераторов вариантов на двоякость и вариативность. В рамках моделирования были рассмотрены онтологии, содержащие классификации генов (GeneOntology [2]) и болезней (ICD-10 [3]). Были посчитаны двоякость (рисунок 1) и вариативность (рисунок 2) терминов в данных онтологий с и без использования генераторов вариантов.

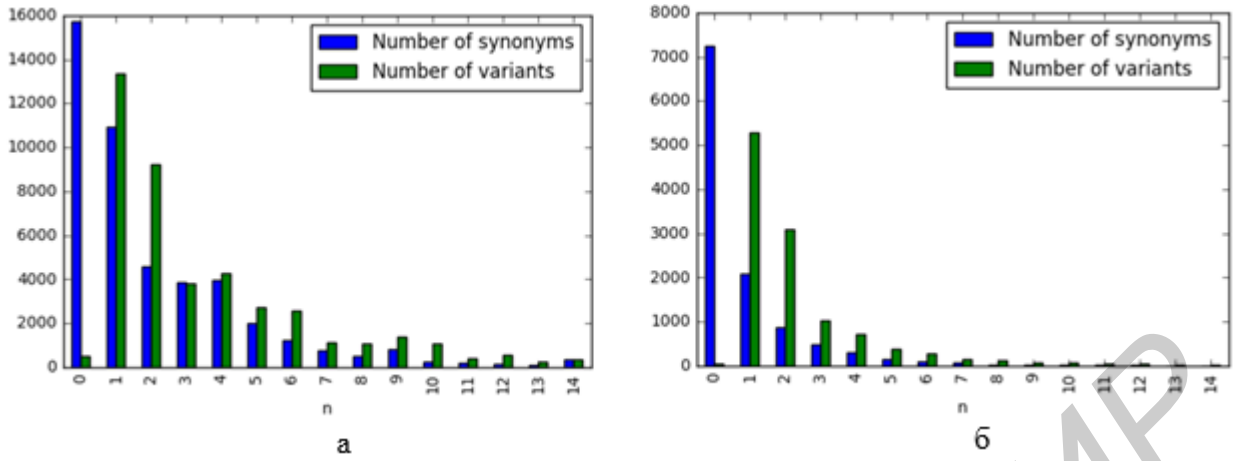

Рис. 1. Количества вариантов написания терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

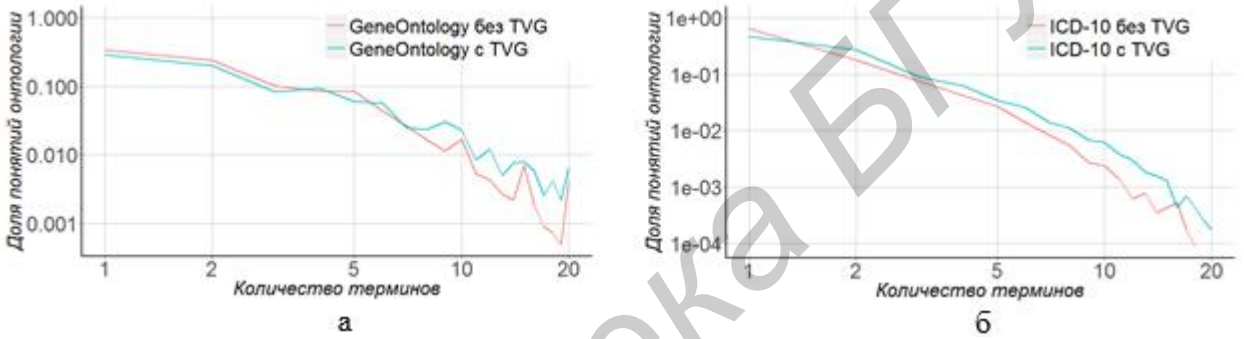

Рис. 2. Изменение двоякости терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

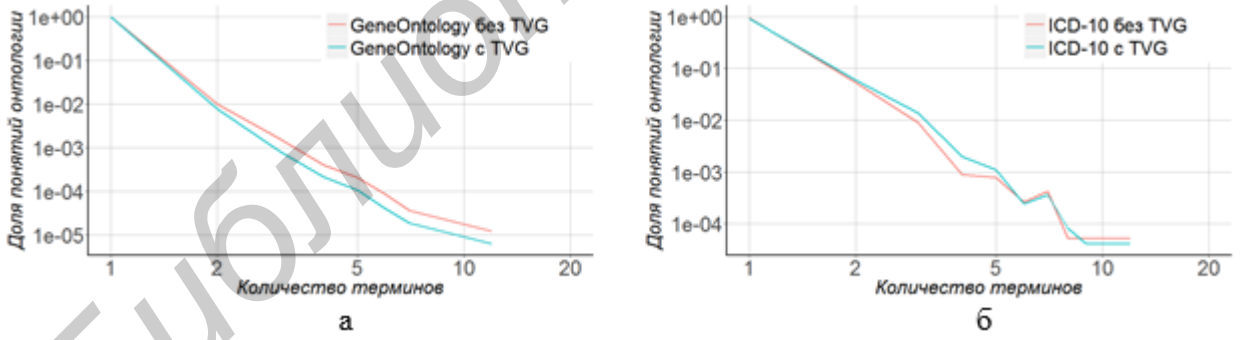

Рис. 3. Изменение вариативности терминов в онтологиях GeneOntology (а) и ICD-10 (б) с и без использования TVG

Из рисунков 2-3 видно, что внедрение генератора не оказало существенного влияния на данные метрики. В то же время, внедрение позволило увеличить полноту поиска (в среднем около 24%), при этом практически не уменьшив его точность (1-2%). Эксперименты показывают, что использование дополнительных вариантов терминов значительно увеличивает шум (количество корректных терминов, для которых найдены некорректные совпадения в онтологиях при разметке). Данную проблему можно решить с помощью внедрения дополнительных фильтров, учитывающих контекст, в котором встречается термин, что является следующим этапом исследования.

Литература

[1]. 1. PubMed NCBI // National Center for Biotechnology Information [Electronic resource]. - 2017. - Mode of access: https://www.ncbi.nlm.nih.gov/pubmed. - Date of access: 13.03.2017.

[2]. 2. Tsuruoka, Y. Normalizing biomedical terms by minimizing ambiguity and variability / Y. Tsuruoka, J. McNaught, S. Ananiadou // BMC Bioinformatics 2008, 9 (Suppl 3) [Electronic resource]. - Mode of access: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-S3-S2. - Date of access: 15.03.2017.

[3]. 2. International Statistical Classification of Diseases and Related Health Problems 10th Revision // World Health Organization [Electronic resource]. - 2016. - Mode of access: http://apps.who.int/classifications/icd10/browse/2010/en. - Date of access: 15.03.2017.

[4]. 3. Gene Ontology Consortium // Gene Ontology Consortium (GOC) [Electronic resource]. - 2015. - Mode of access: http://www.geneontology.org. - Date of access: 14.03.2017.

[5]. 4. Tsuruoka, Y. Probabilistic term variant generator for biomedical terms / Y. Tsuruoka, J. Tsujii: SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval [Electronic resource]. - Mode of access: http://www.nactem.ac.uk/tsuruoka/papers/sigir03.pdf. - Date of access: 15.03.2017.

[6]. 5. Sodium Chloride // Pubchem: Open Chemistry Database [Electronic resource]. - 2017. - Mode of access: https://pubchem.ncbi.nlm.nih.gov/compound/sodium_chloride#section=MeSH-Synonyms. - Date of access: 13.03.2017.