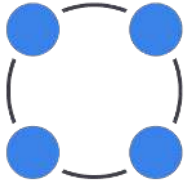
Институт биоинформатики



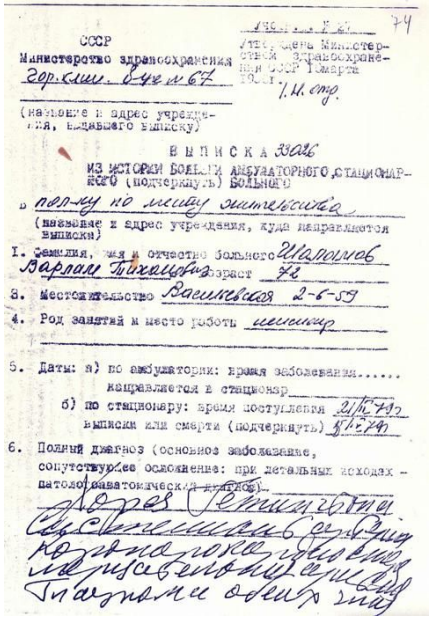
Improving NLP in molecular biology

Руководитель:
Илья Корвиго
МФТИ
Лаборатория функционального анализа генома

Студенты:
Анатолий Зайковский
Максим Холматов

Глобальная проблема:

Формализация биологических данных, написанных на естественных языках, для последующего извлечения из них полезной информации.



СССР
Министерство здравоохранения
Горький. 64г № 67

ВЫПИСКА 3826
ИЗ ИСТОРИИ БОЛЕЗНИ АМБУЛАТОРНОГО, СТАЦИОНАРНОГО (подчеркнуть) БОЛЬНОГО

1. Фамилия, имя и отчество больного Шаломов Варлам Тихонович
2. Местожительство Васильевская 2-6-59
3. Род занятий и место работы инженер
4. Даты: а) по амбулатории: время заболевания.....
   направляется в стационар
   б) по стационару: время поступления 21/IX 73г.
   выписки или смерти (подчеркнуть) 4/X 73г.
5. Полный диагноз (основное заболевание, сопутствующее осложнение: при летальных исходах – патолого-анатомический диагноз)

RESEARCH Open Access

CrossMark

A genetic algorithm enabled ensemble for unsupervised medical term extraction from clinical letters

Wei Liu[1*], Bo Chuen Chung[1], Rui Wang[1], Jonathon Ng[2] and Nigel Morlet[2]

Abstract

Despite the rapid global movement towards electronic health records, clinical letters written in unstructured natural languages are still the preferred form of inter-practitioner communication about patients. These letters, when archived over a long period of time, provide invaluable longitudinal clinical details on individual and populations of patients. In this paper we present three unsupervised approaches, sequential pattern mining (PrefixSpan); frequency linguistic based *C-Value*; and keyphrase extraction from co-occurrence graphs (TextRank), to automatically extract single and multi-word medical terms without domain-specific knowledge. Because each of the three approaches focuses on different aspects of the language feature space, we propose a genetic algorithm to learn the best parameters of linearly integrating the three extractors for optimal performance against domain expert annotations. Around 30,000 clinical letters sent over the past decade from ophthalmology specialists to general practitioners at an eye clinic are anonymised as the corpus to evaluate the effectiveness of the ensemble against individual extractors. With minimal annotation, the ensemble achieves an average F-measure of 65.65 % when considering only complex medical terms, and a F-measure of 72.47 % if we take single word terms (i.e. unigrams) into consideration, markedly better than the three term extraction techniques when used alone.

Keywords: Clinical term extraction, Sequence mining algorithms , Genetic algorithm

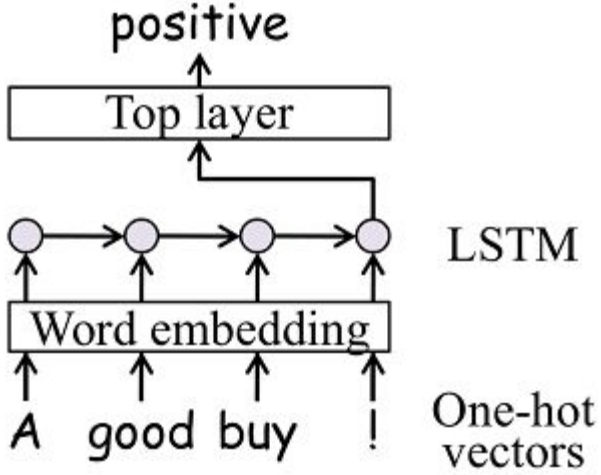
Локальные задачи:

- Ознакомление с литературой по заданной тематике
- Знакомство с современными подходами к анализу естественных языков (NLP)
- Построение классификатора для определения названий химических соединений в форме IUPAC и SMILES в тексте(Анатолий)
- Научится извлекать осмысленные биграммы из текста и оценивать их информативность (Максим)

Word embedding in NLP

Принцип работы:

1. Каждое слово переводится в численный вектор
2. Векторы передаются на слои нейронной сети
3. Выдача результата в виде вектора чисел
4. Обработка результатов



Word embedding and molecular biology

Разновидности:

1. Сопоставив уникальный номер каждому символу из словаря, получаем по слову вектор — последовательность чисел.
2. “One-hot” кодирование: вместо номера каждому символу сопоставляется вектор из нулей и одной единицы, которая стоит на позиции соответствующей номеру. Так слову сопоставляется двумерная матрица из 0 и 1.

Особенности данных молекулярной биологии:

- Очень много уникальных терминов, из-за чего словарь слов становится слишком большим. Значительную часть терминов составляют химические термины.

Задача

Построить классификатор, различающий названия химических соединений, записанных в форме IUPAC и SMILES, и другие слова.

Решение

Рекуррентная нейронная сеть long short-term memory(LSTM).

Данные для тренировки сети

- База названий химических соединений:
  - PubChem

- База обычных слов:
  - Project Gutenberg: математические тексты в формате .tex
  - Художественная литература

Реализация LSTM

- Использование Lasagne — библиотеки python, помогающей работать с библиотекой Theano.
- Сеть обучалась на 120 000 слов, по 40 000 слов из каждой группы: IUPAC названия, SMILES названия и обычные слова.
- Тестовая выборка составляла 10 000 слов.

Результаты

98% точности

Поиск информативных биграмм

Loss of integrin-mediated **cell adhesion**.

Alters receptor specificity, so that transcription is activated by the antiandrogen **cyproterone acetate**.

Substantial reduction of phosphorylation at T-308 and S-473, reduced AKT activation, and reduced binding to PIP3 as well as IGF1-induced **membrane recruitment**. Loss of **membrane localization**; when associated with K-17.

Moderately reduces Ptch1 binding **in vitro** and signaling potency in chicken embryo **neural plate** explant assays compared with wild-type sequence.

Формализация проблемы

- Строим эмбеддинг слов (токенов) в n-мерное векторное пространство.
- Каждому слову соответствует эмпирическое распределение координат векторов-эмбеддингов в контексте этого слова по каждому из n измерений.
- При разбиении биграммы, мы аппроксимируем ее контекст с учетом порядка слов контекстом без учета порядка слов.
- Потеря информации при данной аппроксимации выражается через KL-дивергенцию (Kullback–Leibler divergence).
- Мы хотим терять как можно меньше информации, но минимизировать количество токенов.

Использование частоты колокализации слов:

(('wound', 'healing'), 484)

(('reading', 'frame'), 483)

(('oxide', 'synthase'), 482)

(('wistar', 'rats'), 481)

(('membrane', 'proteins'), 480)

(('data', 'show'), 479)

(('fusion', 'protein'), 475)

(('healthy', 'subjects'), 473)

(('even', 'though'), 471)

(('high', 'affinity'), 471)

(('data', 'demonstrate'), 470)

(('randomly', 'assigned'), 470)

(('fragment', 'length'), 469)

(('protective', 'effect'), 467)

(('study', 'suggests'), 467)

(('membrane', 'protein'), 466)

(('studies', 'showed'), 466)

(('factors', 'associated'), 465)

(('significant', 'effect'), 465)

(('receptor', 'antagonist'), 464)

(('increased', 'significantly'), 463)

(('cell', 'division'), 461)

(('clinical', 'studies'), 461)

(('patients', 'received'), 461)

(('predictive', 'value'), 460)

(('clinical', 'features'), 459)

(('expression', 'profiles'), 459)

(('membrane', 'potential'), 458)

(('study', 'demonstrates'), 458)

(('molecular', 'biology'), 456)

(('mycobacterium', 'tuberculosis'), 455)

(('previously', 'described'), 455)

(('studies', 'using'), 455)

Варианты решения:

- Network centrality measures (degree, betweenness + PageRank, TextRank)
- Статистические подходы:
  - C-value
  - **tf-idf**
- Аппроксимация контекстов Байесовским выводом

\+ комбинации разных подходов

NetworkX

PageRank, TextRank

$$S(v_i) = (1-d) + d \times \sum_{j \in in(v_i)} \frac{1}{|out(v_j)|} S(v_j)$$

Letter 1

He does in fact achieve barely 6/12 unaided, but this improves to 6/6 in each eye separately with a hypermetropic correction. Biomicroscopy showed some nuclear sclerosis in the lens which are quite clear for his age. His intraocular pressures were normal and optic discs and fundi appeared healthy.

Letter 2

Fortunately he still shows no sign of diabetic retinopathy, but is starting to show cataract changes in both eyes even thought this has not affected his sight adversely. His own glasses gave him right 6/9+ left 6/6 and his intraocular pressures were well within normal limits.

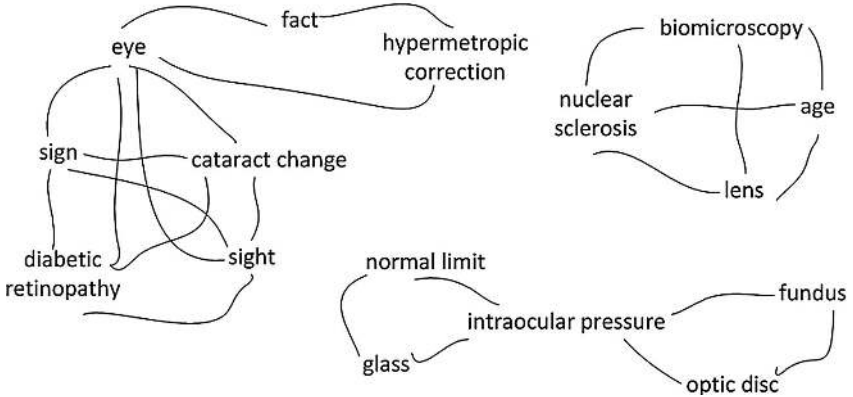


Fig. 1 TextRank example graph. The graph is created by concatenating two clinical documents. Two terms are connected if they appear in the same sentence

Словарь терминов для данного корпуса

('breast', 'cancer') 9.01551716459456

('stem', 'cells') 8.54847683319583

('gene', 'expression') 7.8358973251053925

('cancer', 'cells') 7.020485818052685

('tumor', 'cells') 6.417290970874348

('prostate', 'cancer') 4.749339007303739

('lung', 'cancer') 3.53434663665568

('cell', 'lines') 3.3381303326701706

('cell', 'death') 3.031188273890981

('endothelial', 'cells') 2.9639820342822714

('cancer', 'cell') 2.8537561676368717

('cancer', 'patients') 2.8502312046798846

('gastric', 'cancer') 2.730314488851479

('dna', 'damage') 2.6894924019380677

('epithelial', 'cells') 2.673747879648778

('amino', 'acid') 2.6002360390166745

('dna', 'methylation') 2.4895399978675354

('health', 'care') 2.4895239468295243

('stem', 'cell') 2.454341633681069

('risk', 'factors') 2.3943555481637073

('pancreatic', 'cancer') 2.3879712041787715

('membrane', 'proteins') 2.1903770631203

TF-IDF (term frequency–inverse document frequency)

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

$$\text{tf}(t, d) = \frac{n_t}{\sum_k n_k}$$

$$\text{idf}(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}$$

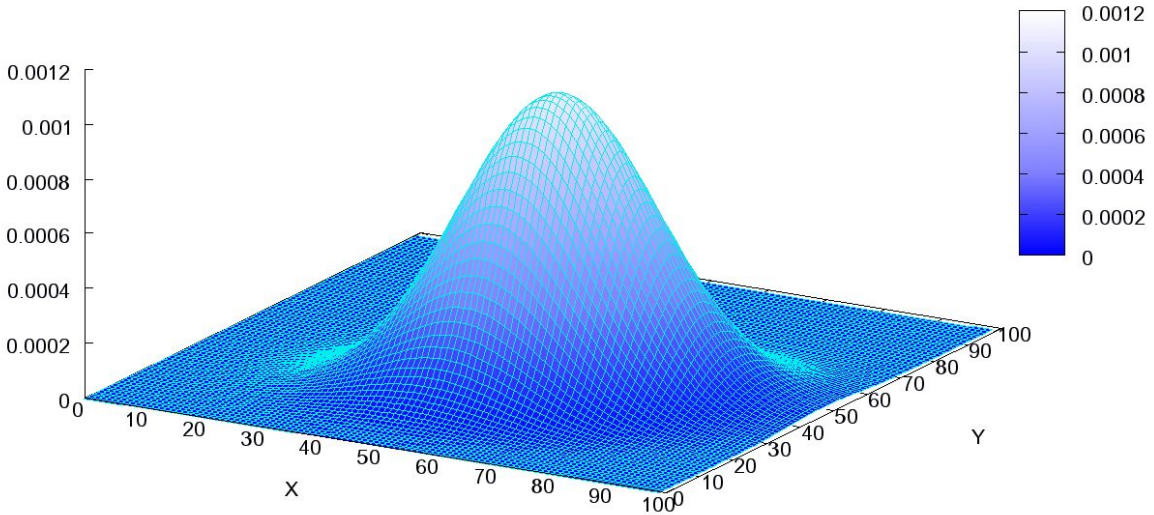
hmg coa : 0.824155060319
ahr ahr : 0.791519318562
ch mab7f9 : 0.788166904694
chick cntf : 0.78103030897
gii gii : 0.777067114578
mir 370 : 0.768058198856
beta tropomyosin : 0.767964492113
facial averageness : 0.765723073784
mir 106a : 0.764482843311
b7 h4 : 0.763269275549
alpha iv : 0.761275993497
mef 2a : 0.760197393292

cmrf 35 : 0.758018483286
la igg : 0.756345395187
cis bf : 0.755574782779
big et : 0.754241074091
alpha 2m : 0.752825616616
danhong injection : 0.750650532189
tip b1 : 0.749102776897
mir 124 : 0.747070338087
trans sas : 0.743867736374
non ar : 0.743861999739
tq nlc : 0.743386752914
ptp pest : 0.742042619431
cuc associated : 0.741605884633

mir 384 : 0.740955698084
ledgf p75 : 0.740812290642
pol beta : 0.740390027896
be1 cells : 0.739881676398
ril beta : 0.738175900044
apoa ii : 0.737340342068
mat alpha : 0.737019131981
extracorporeal treatment : 0.73683803981
cel iii : 0.735550415373
ca cdi : 0.734347649263
tp receptor : 0.734208899709
hip 55 : 0.733159253375
whi 131 : 0.732890105905

Оценка KL-дивергенции

- Для эмпирического распределения сложность экспоненциально зависит от размера эмбеддинга.
- Требуется распределение, для которого можно аналитически найти интеграл любого порядка.
- **Многомерное Гауссово распределение с диагональной матрицей ковариации!**



Спасибо за внимание!