Contents

1 Основные этапы выделения сценария

Выделение сценария состоит из следующих этапов: сегментация и формирование финального сценария на основе полученных сегментов.

Во время этапа сегментации мы обрабатываем каждый текст независимо. Мы предполагаем, что после этого этапа из каждого текста мы получим набор сегментов (отрывков текста из исходного), каждый из которых содержит одну некоторую оформленную мысль и является воплощением одного из шагов сценария. Заметим, что мы можем рассматривать полученную последовательность сегментов, как сценарий без ветвлений, где каждый сегмент - это шаг сценария. Сегментация текста разобрана в сецкии Сегментация.

Формирование сценария может проходить двумя путями:

1. У нас уже есть сценарий и для каждого шага сценария есть примеры сегментов. В этом случае ставится задача не формирования сценария, а его дополнения. В этой задаче на входе у нас есть сегменты текстов, которые должны быть классифицированы по шагам. Это классическая задача классификации и решению этой задачи посвящены разделы 3 и 4.

2. У нас есть только сегментированные тексты и по ним мы должны построить сценарий.

Теперь разберем каждый из этих этапов.

2 Сегментация

На входе мы имеем несколько тесктов, в которых содержатся рекомендации, советы и т.д. для достижения некоторой единой для всех текстов цели. После этапа сегментации мы хотим получить набор отрывков текстов, каждый из которых содержит один шаг сценария.

В данной работе мы предполагаем, что все советы прописаны достаточно явно. В частности, мы предполагаем, что нам не нужно выделять шаги сценария из какого-либо рассказа, в котором повествуется о том, как какой-то человек продвигался к проставленной цели.

2.1 Этапы сегментации

В этоп подразделе разберем основные этапы сегментации. Эту процедуру можно разделить условно на четыре этапа:

1. Разбиение текста на цельные смысловые единницы.
2. Среди смысловых единиц выделить "наиболее важные", центры будущих сегментов.
3. Присоединить оставшиеся смысловые единицы к центрам, используя семантическую близость.
4. Объединить очень близкие сегменты, полученные на прошлом этапе.

Теперь обсудим каждый из этих этапов.

Под цельными смысловыми единицами мы подразумеваем предложения в тексте и списки. Под списками мы подразумеваем некоторые однородные члены предложения, выделенные в тексте таким образом, что каждый член находится на новой строчке, и перед ним стоит некоторый маркер списка. Примерами таких маркеров является число или буква для нумерованного списка или некоторый символ для ненумерованного списка. Понятно, что каждый элемент этого списка не несет особо ценной информации, поскольку это отрывки некоторого предложения. Поэтому эти списки в совокупности с предложением, частью которых они являются, рассматриваются, как единый элемент.

Обратим внимание на то, что мы не рассматриваем списки, элементы которых есть предложения или абзацы. Эти списки так же будут объединены некоторой мыслью, однако в таком случае мы можем получить слишком большие сегменты.

Разбиение на цельные смысловые элементы происходит тривиальным образом, основываясь на пунктуации и символах переноса строки.

Центрами сегментов мы называем предложения, содержащие глаголы или глагольные конструкции такие, что они выражают призыв или рекомендацию к совершению действия. К таким формам мы отнесли

- Глаголы в повелительном наклонении (сделайте),
- Инфинитив с категориями состояния (можно сделать, нужно сделать), причем категории состояния не должны быть зависимы от какого-либо другого слова в предложении,
- Инфинитив с глаголом в третьем лице, настоящем времени без подлежащего (сдедует сделать).

Эти предложения действительно содержат информацию о действии, необходимом для сценария. Однако использовать только эти центры для создания сценария нельзя по двум причинам. Во-первых, как показали эксперименты, при рассмотрении множество только таких центров соответствующие им векторные представления слабо разделимы, т.е. наблюбается крайне плохое качество классификации и не интерпретируемая кластеризация. Во-вторых, часть информации о действии может содержаться в контексте этих центров, что и объясняет первый недостаток. Поэтому мы переходим к следующему этапу сегментации - присоединение к центрам контекста.

Для этого этапа каждому элементу полученному на первом этапе сегментации ставим в соответствие некоторое векторное представление. В наших экспериментах это представление мы получали следующим образом. Каждому слово мы ставили в соответсвие вектор, используя готовые модели word2vec из RusVectores (нужна ссылка), а вектор для элемента текста, находили, как среднее арифметическое векторов для слов, входящих в этот элемент, кроме стоп-слов. Подобное векторное представление достаточно наивное, однако большинство элементов представляют собой достаточно короткие тексты. Как было показано в работе (нужна ссылка), для коротких текстов данный подход приемлим. Таким образом каждому элементу мы поставили с соответствие вектор. Далее мы будем отождествлять понятия элемент текста и вектор.

Все элементы перед первым центром мы относим к первому центру, все элементы после последнего - к последнему. Остальные элементы находятся между двумя центрами. Далее для всех пар соседних центров мы находим разбиение предложений между ними на два непересекающихся множества: контекст первого центра и контекст второго центра. Достаточно естественно предположить, что элементы различных контекстов не должны чередоваться. Поэтому мы ставим дополнительное условие, что разбиение должно быть таким, что все элементы второго множества в тексте находятся после первого множества.

Теперь сформулируем, как мы находим это разбиение. Пусть $\mathbf{c}_1, \mathbf{c}_2$ - первый и второй центр в паре, S - упорядоченное согласно порядку в тексте множество элементов между этими центрами, i - индекс разбиения: все элементы с номером в S меньшим, чем i, относятся к первому множеству, с не меньшим ко второму - i (считаем, что в S нумерация от 0 до $|S|$). Наша задача найти i. Мы будем искать решение следующей задачи:

$$\sum_{j:j<i} d(S_j, \mathbf{c}_1) + \sum_{j:j\geq i} d(S_j, \mathbf{c}_1) \to \min_{i=\overline{0,|S|}},$$

где d - это функция от двух векторов, играяющая роль расстояния. В наших экспериментых мы использовали квадрат евклидового расстояния:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i - y_i)^2.$$

Данная задача решается простым перебором за линейное время от количества предложений между центрами.

В результате последнего этапа мы получаем достаточно полные сегменты с исчерпывающей информацией о действии. Однако, может быть такое, что соседние сегменты есть одно и то же действие. Это может происходить, если автор повторяет одно и то же предложение возможно с небольшими изменениями для обращения внимания читателя на какой-то важный момент. Поэтому некоторые сегменты следует так же объединить. Во-первых, это уменьшит количество сегментов, которые нежно будет обрабатывать в будущем. А во-вторых, это уточнит сегменты, увеличив вероятность их правильной классификации.

Мы получаем векторные представления для полученных сегментам аналогично векторным представлениям для элементов, что описано выше. Для каждой пары соседних сегментов высчитываем расстояние $d'(\mathbf{c}_1, \mathbf{c}_2)$ между ними и объединяем группу соседствующих сегментов, если расстояние между парами соседних сегментов в этой группе меньше некоторого порога ϵ. К примеру, допустим, что у нас есть последовательность сегментов $\mathbf{c}_k, k = \overline{1,5}$, таких что

$$d'(\mathbf{c}_1, \mathbf{c}_2) < \epsilon; d'(\mathbf{c}_2, \mathbf{c}_3) < \epsilon;$$

$$d'(\mathbf{c}_3, \mathbf{c}_4) > \epsilon; d'(\mathbf{c}_4, \mathbf{c}_5) < \epsilon.$$

В таком случае сегменты 1-3 будут объединены в один сегмент, а сегменты 4 и 5 в другой.

В качестве расстояния мы используем следующую функцию:

$$d'(\mathbf{c}_1, \mathbf{c}_2) = \text{WMdistance}(\mathbf{c}_1, \mathbf{c}_2) + \alpha\sigma(l_1 + l_2),$$

где WMdistance - Word Mover's Distance (см. нужна ссылка), $\sigma(x) = \frac{1}{1+e^{-x}}$ сигмоид, монотонно возрастающая ограниченная функция функция,

l_1, l_2 - длины первого и второго сегмента соответственно, α - параметр, отвечающий за величину штрафа. Использование WMdistance обозначено тем, что для сегментов, часть из которых, уже стали достаточно большими текстами, он показал лучшие результаты, чем остальные метрики. Однако, мы предполагаем, что короткие сегменты скорей всего не представляют сами по себе большой ценности. Поэтому мы используем сигмоид, который поощряет объединение коротких и штрафует объединение длинных сегментов, причем штраф практически не различается для очень длинных длинных текстов.

3 Об идентификации шагов

В данном разделе мы хотим изучить, насколько хорошо возможно отличить один шаг от другого и насколько хорошо можно определить, что два шага из разных документов есть один и тот же шаг.

3.1 Гипотеза О Корректной Идентификации Шагов

Мы предполагаем, что каждому шагу соответствует некоторый вектор и вектора текстов, которые воплощают один и тот же шаг, ближе всего к центрального вектора именно этого шага. Если эта гипотеза выполняется, то шаги действительно хорошо идентифицируются. Эту гипотезу мы назовем гипотезов о корректной идентификации и её проверки и будет посвущен этот раздел.

Соответствие между текстом и вектором строится следующим образом: возьмем некоторую обученную модель word2vec (в нашем случае мы использовали готовые модели RusVectores), каждому слову в тексте поставим в соответствие вектор из этой модели, и тогда вектор, соответствующий этому тексту, есть среднее арифметическое векторов всех слов входящих в него.

Мы проведем следующий синтетический эксперимент для проверки нашей гипотезы:

1. Каждый текст вручную разбиваем на сегменты, каждый из которых есть действие относящиеся к некоторому шагу;

2. Вручную делаем разметку сегментов;

2. При помощи word2vec для каждого сегмента мы находим соответствующий ему вектор;

3. Для каждого шага сценария мы примем за центр среднее арифметическое соответствующих векторов;

4. Определим для каждого сегмента шаг по ближайшему центру и сравним полученную разметку с исходной разметкой.

Информацию о том, какие шаги были выделены и сколько текстов для каждого шага в нашем наборе, Вы можете найти в таблице 1.

№ шага	Имя шага	Количество текстов
0	Ваши деньги	5
1	Цены	4
2	Объявлеия	4
3	Телефонный разговор	7
4	Документы на машину	6
5	Мониторинг Сайтов	9
6	ДКП	8
7	Осмотр	8
8	Тест-драйв	5
9	Определиться с маркой и моделью	7
10	Дигностика	4
11	Год выпуска и пробег	3

Table 1: Информация о выделенных шагах

Для оценки качества распознавания каждого шага мы будем использовать две метрики: Precision и Recall (см. таблица 2). Стоит заметить, что некоторые шаги распознаются значительно хуже (к примеру Recall четвертого шага составляет всего 66.7%, т.е. около трети текстов, относящихся к этому шагу не было распознано). Это связано с двумя факторами. Во-первых, шаги не являются сильно детализированнами. Это приводит к тому, что, к примеру, в разделе 'Телефонный разговор' могут быть советы, связанные к документации, что приближает такие тексты к пятому шагу. А во-вторых, некоторые шаги близки по смыслу сами по себе. К примеру, очевидно, что близкими являются пары 4-6 и 7-8. Первый фактор устраняется достаточно легко - большая детализация. А ухудшения качества, вызванных вторым фактором, пожалуй можно и не считать сильных ухудшением. Ведь если шаги достаточно близки, то небольшая путаница в текстах, скорей всего, не повлияет на качество финального результата.

№ шага	Recall,%	Precision, %
0	100.0	100.0
1	100.0	100.0
2	100.0	80.0
3	85.7	66.7
4	66.7	75.0
5	83.3	83.3
6	75.0	85.7
7	87.5	87.5
8	80.0	100.0
9	100.0	100.0
10	100.0	100.0
11	100.0	100.0

Table 2: Качество идентификации шагов

Для глобальной оценки идентификации шагов, мы используем усредненные Precision и Recall, а так же Accuracy (см. таблица 3). Заметим, что выборка является достаточно сбалансированной - размер всех шагов, кроме последнего, отличается от среднего не более, чем в полтора раза. Мы получили достаточно высокие показатели - порядка 90%.

Метрика	Значение, %
Mean Recall	89.9
Mean Precision	89.9
Accuracy	87.1

Table 3: Качество идентификации шагов, средние показания

На основе полученных выше результатов, можно считать гипотезу об корректной идентификации шагов выполненной с высокой точностью. Далее обсудим растояние между шагами и зависимость качества разделяемости шагов от количества текстов, на основе которых строится центральный вектор.

3.2 Влияние Размера Выборки Тестов

Выше описанные эксперименты и результаты говорят о том, насколько хорошо возможно идентифицировать шаги. Рассмотрим более реальную

задачу:

- На входе n текстов, из которых для $m \leq n$ нам известен номер шага к которому они принадлежат.
- На выходе разметка для всех n шагов.

В данном разделе изучим, как зависит от качество результата от m. Для этого мы сделаем следущее:

- Оставим из исходной выборки, описанной в таблице 1, только 4 шага и тексты, относящиеся только к ним. В нашем случае мы оставили шаги 4, 6, 7, 9.
- Сделаем так, чтобы для каждого шага было только k текстов. В нашем случае мы взяли $k = 7$.

Выделим из полученных данных обучающую (та часть, на основе которой мы будем строить центральные вектора для шагов) выборку: выберем r текстов для каждого шага. Таким образом r - это количество текстов для каждого шага. Сделаем всемозможные выборки для фиксированного r, решим выше сформулированную задачу для них, измерим качество, усредним метрики и полученные значение метрик примем за метрики качества для r текстов на шаг. Полученные результаты представлены в таблице 4. Значения Recall и Accuracy достаточно близки поэтому на графике 3.2 находится данные только для метрик Recall и Precision.

r	Recall, %	Precision, %	Accuracy, %
1	63.4	73.1	63.4
2	75.8	79.7	75.8
3	83.5	85.5	83.5
4	88.6	89.8	88.6
5	92.2	93.0	92.2
6	95.2	95.7	95.2
7	96.4	96.9	96.4

Table 4: Влияние Размера Выборки Тестов На Качество Идентификации

мы получаем следующие результаты. Во-первых, увеличение размера исходной разметки ведет к улучшению качества по любой метрике. Данный

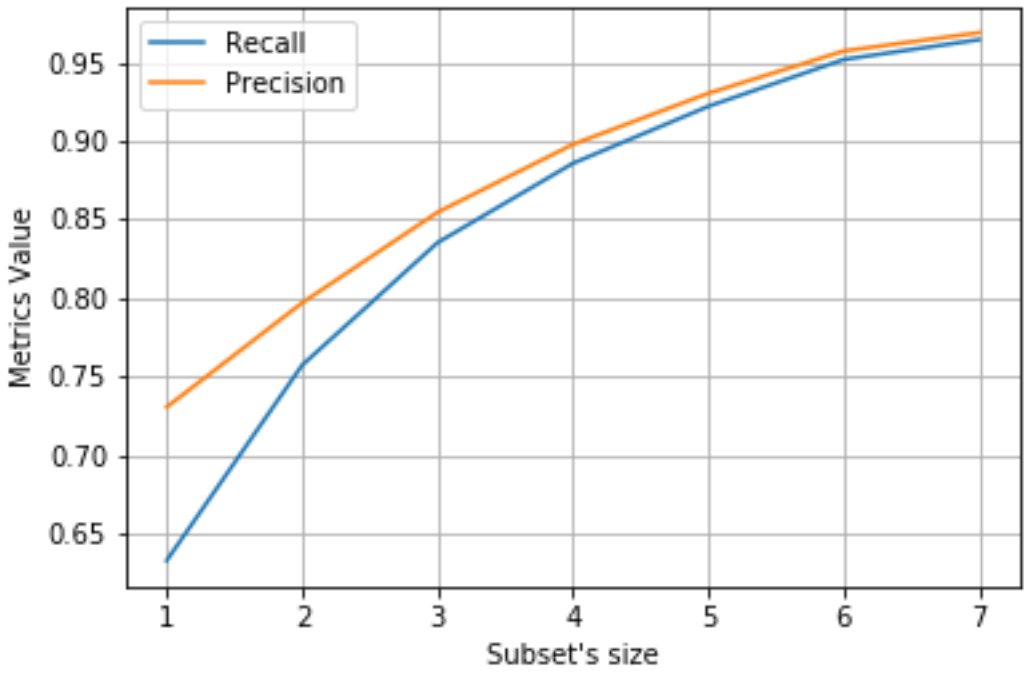


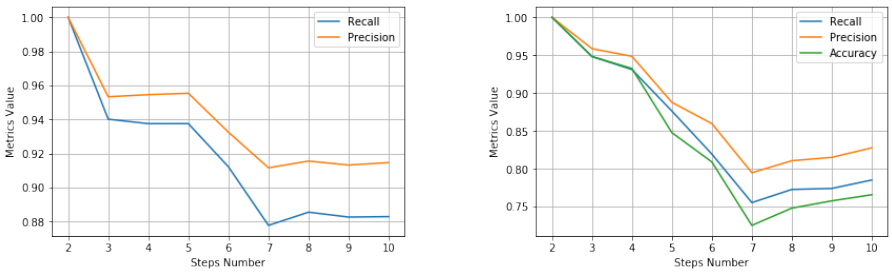
Figure 1: Зависимость качества от размера выборки

момент был достаточно очевиден и без данного эксперимента, поскольку это есть следствие того, что чем больше текстов для каждого шага дано, тем точнее мы определим центральный вектор этого шага. Во-вторых, при шести текстах достигается качество 95% и далее активный рост прекращается. Таким образом, шесть текстов на шаг, когда всего четыре шага, можно считать необходимым и достаточным количеством для качественной идентификации. Теперь зададимся вопросом, насколько этот показатель качества ухудшится, если количество шагов продолжит расти.

3.3 Влияние Количества Шагов

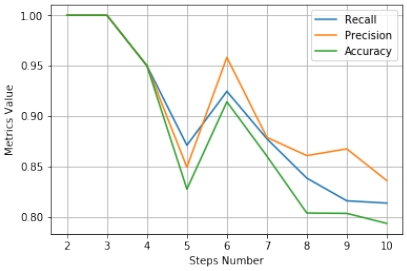
Проведем серию экспериментов, измеряя качество на каждом этапе:

- **3/4**: Каждый шаг содержит 4 текста, 3 из них обучающие. Количество



(a)

(b)



(c)

Figure 2: Влияние количества шагов на качество распознавания при фиксированном размере обучающей выборки: (a) 3/4; (b) 3/ALL; (c) 4/ALL.

варьируется шагов от 2 до 11.

- **3/ALL**: Каждый шаг содержит столько текстов, сколько было в исходном, 3 из них обучающие. Количество шагов варьируется от 2 до 11.

- **4/ALL**: Каждый шаг содержит столько текстов, сколько было в исходном, 4 из них обучающие. Количество шагов варьируется от 2 до 11.

Результаты этих экспериментов представлены на графиках 2.

Из полученных результатов мы можем сделать следующие выводы:

1. Качество идентификации падает при увеличении количества шагов в начале. Этот вывод достаточно очевиден и ожидаем.

2. Падение качества замедляется после семи-восьми шагов.

3. Падение качества останавливается приблизительно на значении 80%.

Отсюда следует, что можно ожидать, что шесть текстов на шаг (количество, установленное в предыдущем подразделе для семи шагов) будет и далее показывать хорошие результаты.

4 Классификация

В предыдущем разделе мы показали, что выделенные человеком сегменты текста можно классифицировать с достаточно высокой точностью, имея разметку для достаточно небольшого количества сегментов. В данном разделе мы изучим следующую задачу классификации:

- Объекты - программно выделенные сегменты класса(см. **??**),
- Признаки - их векторное представление,
- Метка класса - номер шага.

4.1 Данные

В данном разделе мы посмотрим, что из себя представляют наши данные в выбранном представлении. По сравнению с исходными классами мы добавили 12 класс - это класс бесполезных, мусорных сегментов. На данном этапе мы их не будем рассматривать ни при обучении, ни при тестах.

Как можно видеть на рис. 3, выборка достаточно несбалансированна. Так седьмой шаг представлен является самым большим по количеству экземпляров и составляет больше половины всех имеющихся данных.

Так же на рисунке 4 можно заметить, что данные являются достаточно перемешанными. Что означает, что корректно классифицировать шаги будет нетривиальной задачей.

4.2 Наивная классификация

Наивная классификация заключается в использовании той же модели, что мы использовали в предыдущем разделе. Каждому сегменту ставится в соотвествие вектор, ка среднее арифметическое векторов предложений,

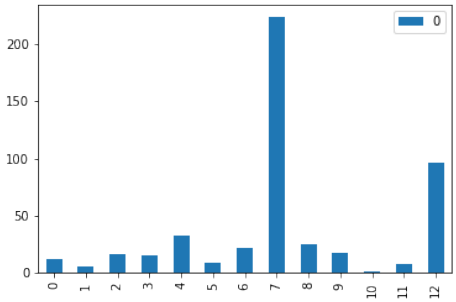


Figure 3: Распределение данных

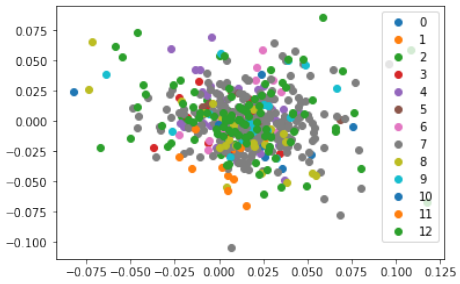


Figure 4: Визуализация

которые в свою очередь есть среднее векторов слов. По размеченным данным мы получаем центр каждого класса, как среднее арифметическое соответствующих помеченных векторов. Для новых объектов мы определяем класс по ближайшему вектору. Измерим качество классификации на кросс=вадидации (1000 разбиений, размер обучающей выборки - 0.8 от длины всей выборки.) по метрикам Precision и Recall для каждого класса и усредненным метрикам. Результаты Вы можете найти в таблице 5.

Наивная классификация дает результат лучше, чем случайное угадывание (ожидаемое значение точности около 9%). Это говорит о применимости выше описанного векторного представления для сегментов.

Однако большинство классов кроме седьмого и одиннадцтого распознаются достаточно плохо. Более того, большинство текство отнесенных к классам кроме выше названных отнесены ошибочно (метрика Precision). Это является достаточно большой проблемой.

Посмотрим на результаты более сложных классификаторов

4.3 Логистическая регрессия

№ шага	Precision, %	Recall, %
0	20	16
1	0	0
2	37	41
3	17	46
4	43	41
5	2	5
6	34	36
7	87	64
8	28	41
9	42	48
11	78	85
MeanValue	35	39
Accuracy,%	52	

Table 5: Значения метрик на кросс-валидации для наивной классификации