

# РАСПОЗНАВАНИЕ ИНТЕНЦИЙ ПОКУПАТЕЛЕЙ В СООБЩЕНИЯХ СОЦИАЛЬНЫХ СЕТЕЙ (НА ПРИМЕРЕ СЕТИ “В КОНТАКТЕ”)

И. С. Пименов<sup>1</sup>, Н. В. Саломатина<sup>2</sup>

<sup>1</sup> *Новосибирский государственный университет, 630090, Новосибирск*

<sup>2</sup> *Институт математики СО РАН, 630090, Новосибирск*

УДК 81'322.2

В работе рассматривается задача автоматического обнаружения интенций (выраженных явно или неявно намерений совершить некоторое действие) в сообщениях пользователей сети Интернет. В основе предложенного решения — использование шаблонов интенций, построенных экспертом с применением результатов  $N$ -граммного анализа обучающей коллекции текстов (сообщений покупателей сети “В контакте”). Совокупность шаблонов интенций представлена в виде ориентированного ациклического графа, позволяющего группировать схожие маркеры (слова и словосочетания, указывающие на присутствие интенции в тексте) на основании общности грамматических характеристик. Программа поиска интенций реализована на языке Python. Точность и полнота распознавания тестовой коллекции, сформированной по основному и газетному корпусу русского языка, составляет 80% и 74% соответственно.

**Ключевые слова:** интенция, маркер интенции,  $N$ -граммный анализ, ориентированный граф.

## Введение

Значительное количество торговых сделок заключается посредством социальных сетей, которые одновременно характеризуются и высоким уровнем активности пользователей, и отсутствием единой системы организованной регуляции экономических отношений. Оптимизация инфраструктуры виртуального рынка возможна, например, за счёт разработки программных средств, позволяющих связать продавца и покупателя, заинтересованных в сделке по одному и тому же товару или услуге и разделенных информационным шумом. Задача автоматического обнаружения намерений (интенций) покупателей относится к проблемам поиска и извлечения фактов из текстов.

Факт, как некоторая структура, представляет собой связанные определенным отношением сущности. Соответственно, поиск факта в тексте предполагает осуществление процедуры опознавания сущностей и определения связей между ними.

Как правило, различают два основных подхода к решению данной задачи:

- 1) *экспертный*: систему правил (шаблонов) извлечения фактов строит эксперт, опираясь как на собственный опыт, так и на статистики, полученные путем автоматической обработки данных [1, 2],
- 2) *машинное обучение*: правила строятся автоматически на базе обучающих коллекций, размеченных экспертом [3, 4].

В настоящей работе апробирован первый подход.

Все многообразие реализаций фактов в тексте задается правилами, которые могут быть представлены, к примеру, регулярными выражениями, контекстно-свободными грамматиками [5], специальными конструкциями типа шаблонов, фреймов [6] и др. Конструкции на основе шаблонов являются наиболее распространенными.

Автоматизация построения шаблонов достигается различными средствами, к примеру:

- 1) итерационным методом: построенные экспертом шаблоны обогащаются именами сущностей и связями между ними путем поочередного поиска в текстовых коллекциях известных сущностей с неизвестными связями и известных связей с неизвестными сущностями [6],

2) методом с применением синтаксического анализа: сущности и связи извлекаются из дерева разбора предложения [7] и пр.

Поиск клиентов (установление наличия в тексте факта покупательской интенции) в русскоязычных социальных сетях в помощь производителям товаров и услуг предлагает, например, группа Лидсканер [9]. Результаты поиска выдаются в виде списков сообщений отнесенных к определенным тематическим классам, таким как „аренда недвижимости“, „бытовой ремонт“, „медицинские услуги“ и др. Извлечение объекта интенции и идентификация его свойств не предполагается. Поэтому в один класс попадают, например, сообщения о намерениях получить консультацию у врачей различных профилей, работающих в разных городах. Выявление объекта и конкретизация его свойств позволит существенно уточнить классификацию.

Цели исследования:

- 1) по обучающей коллекции текстов, содержащих покупательские интенции, построить правила для установления факта интенции, извлечения из сообщения самого объекта и его свойств;
- 2) в терминах полноты и точности оценить качество полученных правил на тестовой коллекции.

## 1 Задача распознавания интенций

Неформально интенция определяется как выраженный эксплицитно или имплицитно факт желания совершить какое-либо действие, например:

(1) эксплицитно: „хочу устроить фотосессию девушке на этих выходных в Москве“;

(2) имплицитно: „подскажите пожалуйста хорошего репетитора по русскому языку“.

(Во всех примерах используется оригинальный текст сообщений пользователей социальных сетей, он может содержать ошибки.)

Очевидно, что неформальное определение интенции является малопродуктивным для целей поиска и извлечения информации. Его формализация предполагает выявление структуры интенции, а именно: из каких элементов она состоит, какие элементы являются обязательными, какие — факультативными, каков порядок следования элементов в структуре. Необходимо также установить допустимую вариативность элементов на лексическом и грамматическом уровне и возможные изменения в элементном составе структуры на уровне редакционных операций. На основании полученных результатов требуется построить поисковые шаблоны, осуществить их агрегацию согласно общности структуры и свойств, дать оценку их качества на тестовом материале.

## 2 Используемые методы

Экспертное заключение о структуре интенции ( $I$ ) было получено на базе обучающего материала (сообщений пользователей социальных сетей, содержащих интенции). Она состоит из следующих элементов: маркера факта интенции ( $Mr$ ) — слова или словосочетания, указывающего на наличие интенции в тексте (и, как правило, предворяющего ее), объекта интенции ( $O$ ), а также ограниченного числа ( $k$ ) его свойств  $p_i$ , ( $i = 1, \dots, k$ ) и/или конструкций вежливости ( $E$ ). В примере (2) предыдущего раздела маркером интенции является слово „подскажите“, объектом интенции — „репетитора“, свойствами объекта, несущественным — „хорошего“ и существенным — „по русскому языку“, формулой вежливости — „пожалуйста“. К свойствам относится и указатель на объект: *телефон* репетитора. Шаблон интенции в обобщенном виде может быть представлен последовательностью элементов:

$$Pattern(I) = Mr[X_j][E][X_j][p_i][X_j]O[X_j][p_i],$$

где в квадратные скобки заключены факультативные (необязательно присутствующие в сообщении) элементы. Символами  $X_j$  обозначаются возможные (ограниченные по длине) вставки в структуру интенции не всегда семантически значимых единиц текста сообщения (подскажите *мне*, подскажите *кто-нибудь*). Маркер интенции маловариативен, на лексическом уровне его можно считать константой, поэтому в шаблоне он представлен конкретной леммой и грамматическими характеристиками возможных ее форм в тексте. Объект может быть практически любым на уровне лексики, в шаблоне он фиксируется только на грамматическом уровне, на котором маловариативен. Факультативные элементы, как и основные, задаются на лексическом или грамматическом уровне в зависимости от степени изменчивости. Формулы вежливости перечислимы, имеется список их вариантов.

Совокупность шаблонов интенций удобно представлять в виде ориентированного ациклического графа, позволяющего группировать схожие по структуре шаблоны с разными маркерами на основании общности грамматических характеристик (и, как следствие, синтаксической функции) его элементов, а также оптимизировать процедуру поиска, устраняя проблемы частичной заполненности факультативных позиций в шаблоне.

При построении графов в помощь эксперту был проведен  $N$ -граммный анализ обучающей коллекции текстов, содержащих интенции. Под  $N$ -граммой понималась цепочка  $X$  из  $N$  подряд следующих элементов текста (словоформ, лемм, граммем).  $N$ -граммный анализ позволил: 1) дополнить словарь маркеров, построенный экспертом, 2) учесть вариативность структуры интенций.

Процедура  $N$ -граммного анализа включала следующие действия:

1. Нормализацию и морфологический анализ обучающей коллекции текстов.
2. Формирование цепочек  $N$ -грамм с вычислением их абсолютной  $F_a$  и тестовой  $F_t$  частоты встречаемости в коллекции.
3. Фильтрацию  $N$ -грамм по критерию „маркерности“: цепочка является маркером, если выполняется условие:  $(C > P_1) \& (F_t > P_2)$ , где  $C = F_t/F_a$ . Пороги  $P_1$  и  $P_2$  были определены экспериментально. На базе  $N$ -грамм, удовлетворяющих критерию, формировался словарь потенциальных маркеров интенций.
4. Исследование  $N$ -грамм на вариативность, а именно: возможность реализации в любой позиции цепочки замен, ограниченных по длине вставок (см. [10]). Полученные результаты позволили устанавливать для каждой вершины графа вершины, смежные с ней. Кроме того, количественно была охарактеризована степень вариативности элементов цепочки на лексемном и граммежном уровне.

Каждой вершине графа поставлен в соответствие элемент шаблона (информация о лексеме и ее грамматических характеристиках). Процедура распознавания интенции начинается с поиска в нормализованном тексте маркера интенции. Список маркеров зафиксирован в словаре. Обнаружение маркера интенции в тексте определяет выбор графа, его начальную вершину. Дальнейший проход по тексту сопровождается сравнением лексемы и/или грамматической информации о ней с информацией, соответствующей вершине графа. При совпадении осуществляется переход к следующему слову текста и следующей вершине. Иначе — последовательно просматриваются все другие смежные вершины. При отсутствии вершины с подходящей информацией процедура распознавания заканчивается с результатом „интенция отсутствует в тексте“. Считается, что интенция содержится в тексте, если в ходе его анализа найдется хотя бы один граф, из начальной вершины которого может быть достигнута его конечная вершина. Если интенция обнаружена, исходя из информации, соответствующей вершинам графа, легко может быть извлечен объект интенции и его свойства.

## 3 Апробация подхода

### 3.1 Построение графов интенций

Текстовая коллекция, участвовавшая в обучении, содержала 1203 сообщения из социальной сети „В контакте“, которые отбирались экспертом с помощью API LeadScanner [9]. Для нормализации и морфологического анализа текста были использованы модули PyMorphy2 версии 3.5.2 [11].

По результатам  $N$ -граммного анализа в первую версию словаря попали 17 однословных маркеров. Топ упорядоченных по убыванию  $S$ -характеристик  $N$ -грамм, претендентов на маркеры, представлен в таблице 1 ( $P_1 > 0,9$ ;  $P_2 > 85$ ). В первом десятке списка содержится 4 маркера, они выделены курсивом. Остальные слова, хотя и не являются маркерными, но часто им сопутствуют („срочно нужен“, „подскажите, пожалуйста“, ...). При продвижении в конец списка число „шумящих“  $N$ -грамм возрастает.

Исследование  $N$ -грамм на вариативность упростило построение графа в плане установления смежных вершин и инцидентных им ребер. Например, все возможные замены слова, следующего за маркером „подскажите“, позволяют установить потенциальные продолжения: „подсказать *пожалуйста*“ (формула вежливости), „подсказать *хороший* (терапевт)“ (свойство объекта, эмфатический элемент), „подсказать *мануальный* (терапевт)“ (свойство объекта), „подсказать *номер* (терапевта)“ (указатель на объект).

Число вершин в построенных графах равно длине интенции в словах, оно может доходить до полутора десятков. Число различных путей достижения конечной вершины из начальной — до 200. Оно характеризует

Таблица 1: Список  $N$ -грамм, претендентов на маркеры интенций

Лемма	C-характеристика	$F_t$
<i>посоветовать</i>	1,00	119
срочно	0,99	107
<i>подсказать</i>	0,98	278
искать	0,97	211
можно	0,97	120
пожалуйста	0,96	225
хороший	0,96	203
<i>нужный</i>	0,95	162
<i>где</i>	0,93	86

разнообразие способов выражения интенции в тексте. Граф с начальной вершиной, которой соответствует информация о маркерах глагольной группы („подсказать“, „посоветовать“, „порекомендовать“), представлен на рисунке 1. Он имеет 15 вершин (15-я конечная не обозначена для сохранения наглядности). Пять вершин являются смежными с конечной: 2, 7, 9, 10, 12.

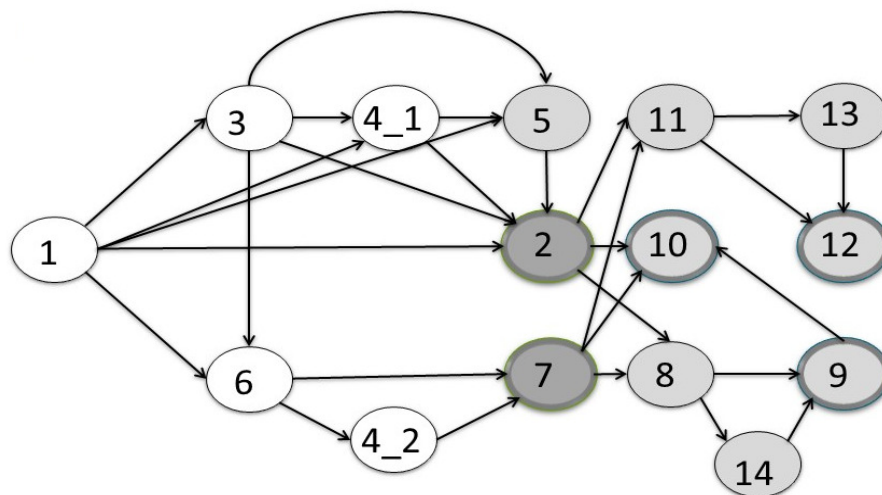


Рис. 1: Пример графа интенций, предваряемых группой глагольных маркеров „подсказать“, „посоветовать“, „порекомендовать“.

Содержание вершин (роли лексем и их грамматические характеристики):

- 1 — глагол, императив, множественное число (маркер: „подсказать“, „посоветовать“, „порекомендовать“, „предложить“);
- 2, 7 — существительное, аккузатив (непосредственный объект интенции);
- 3 — формула вежливости („пожалуйста“ и его сленговые варианты);
- 4\_1, 4\_2 — прилагательное, аккузатив (эмфатический элемент: „хороший“, „толковый“, „качественный“, и др.);
- 5 — прилагательное, аккузатив (свойство объекта интенции: „*мануальный* терапевт“);
- 6 — существительное аккузатив (указатель объекта: „контакты“, „номер“, „телефон“);
- 8 — предлог (свойство объекта интенции: „репетитора *по* физике“);
- 9 — существительное, датив (свойство объекта интенции: „репетитора *по физике*“);
- 10 — существительное, номинатив (указатель места: „репетитора *по ивриту Новосибирск*“);
- 11 — предлог, лексическая константа „в“ (указатель места: „косметолога *в* Сарове“);
- 12 — существительное, локатив (указатель места: „косметолога *в Сарове*“);
- 13 — прилагательное, локатив (указатель места: „косметолога *в нашем* городе“);

14 — прилагательное, локатив (свойство объекта интенции: „массажиста в *коминтерновском* р-не“).

### 3.2 Результаты тестирования графов интенций

Материал для тестовой коллекции был подобран вручную из Национального корпуса русского языка (основного и газетного подкорпусов) [12] с помощью встроенной поисковой системы. Каждый запрос содержал один из маркеров интенции. Тестовую коллекцию составили 110 сообщений. Некоторые из них не включали интенцию. Кроме того, не все интенции относились к категории покупательских (например, „Подскажите пожалуйста длительность курса“).

Для оценки полноты ( $R$ ) и точности ( $P$ ) распознавания применялись меры, которые обычно используются при оценивании алгоритмов поиска и извлечения информации:  $P = TP / (TP + FP)$ ;  $R = TP / (TP + FN)$ , где  $TP$  — количество релевантных ответов (решение эксперта и программы совпадают),  $FN$  — число ошибок первого рода,  $FP$  — число ошибок второго рода. В таблице 2 приведены значения полноты и точности на обучающей ( $L$ ) и тестовой ( $T$ ) коллекции с макроусреднением результатов.

Таблица 2: Полнота и точность распознавания интенций на обучающей и тестовой коллекции

	$L$	$T$
$R$	0,82	0,74
$P$	0,99	0,80

Ошибки распознавания, выявленные на обучающей и тестовой коллекции, связаны с проблемами разных уровней сложности. Самые трудноразрешимые относятся к концептуальным проблемам области, они вызваны принципиальной языковой сложностью выражения интенции как психологической категории. В отличие от них недостаточная детализация в конструкции построенных графов может быть легко устранена. Ошибки морфологического анализатора, орфографические и синтаксические ошибки пользователей социальных сетей являются естественным „шумом“, на который есть смысл настраивать содержимое вершин графа (например, указывать варианты лексем и граммем, не предусмотренные орфографией и синтаксисом корректно написанного сообщения) в силу того, что они носят регулярный характер и не создают трудностей, связанных с омонимией.

Следует отметить, что объект и его свойства извлекались с приемлемым качеством и из интенций, не относящихся к покупательским, что свидетельствует о возможности расширения области применения уже построенных графов.

## Заключение

В данной работе предложен и апробирован подход к распознаванию интенций покупателей в сообщениях социальных сетей, основанный на применении построенных экспертом правил в форме лексико-синтаксических шаблонов, агрегированных в графы. Применение  $N$ -граммного анализа в качестве средства автоматизации позволило дополнить словарь маркеров интенций, упростить построение графов.

Апробация подхода показала, что использованный подход, дает возможность устанавливать, содержится ли интенция в тексте, и, если она содержится, то извлекать объект интенции (а также его свойства) с хорошими показателями полноты и точности вне зависимости от того, имеет ли интенция покупательскую специфику.

## Список литературы

- [1] Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Труды межд. конф. «Диалог 2004»: Компьютерная лингвистика и интеллектуальные технологии. 2004. С. 282–285.
- [2] Хорошевский В.Ф. OntosMiner: семейство систем извлечения информации из мультязычных коллекций документов // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ–2004. Труды конференции в 3-х томах. М.: Физматлит, 2004. Т. 2. С. 573–581.

- [3] Pande V., Khandelwal A.S. Information extraction technique: a review. // National Conference on Recent Threads in Computer Science and Information Technology (NCRTCSIT-2016). IOSR Journal of Computer Engineering (IOSR-JCE).
- [4] R.Navigli, P.Velardi. LearningWord-Class Lattices for Definition and Hypernym Extraction [Электронный ресурс]. <http://www.aclweb.org/anthology/P10-1134> (дата обращения: 23.03.2017).
- [5] Томита-парсер [Электронный ресурс]. <https://tech.yandex.ru/tomita> (дата обращения: 26.04.2017).
- [6] Большакова Е., Баева Н., Бордаченкова Е. и др. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2007. Т. 2. Изд-во РГГУ Москва, 2007. С. 70–75.
- [7] Лукашевич Н.В. Итерационное извлечение шаблонов описания событий по новостным кластерам. // Труды конференции RCDL-2012 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Переславль-Залесский). 2012. С. 353–359.
- [8] RCO Fact Extractor SDK [Электронный ресурс]. [http://www.rco.ru/?page\\_id=3554](http://www.rco.ru/?page_id=3554) (дата обращения: 20.04.2017)
- [9] Leadscanner [Электронный ресурс]. <https://leadscanner.ru> (дата обращения: 12.04.2017).
- [10] Гусев В.Д., Саломатина Н.В. Уточнение и обогащение индикаторных словарей для автоматического извлечения информации из научных текстов. // Труды межд. конф. "Диалог'2007": Компьютерная лингвистика и интеллектуальные технологии, Бекасово, 31 мая–4 июня. 2007. С. 486–491.
- [11] Rymorphy2 [Электронный ресурс]. <https://rymorphy2.readthedocs.io/en/latest/> (дата обращения: 16.02.2017).
- [12] Национальный корпус русского языка [Электронный ресурс]. <http://www.ruscorpora.ru/search-main.html> (дата обращения: 11.03.2017).

*Иван Сергеевич Пименов — студент Новосибирского государственного университета;  
e-mail: pimenov@yandex.ru;*

*Наталья Васильевна Саломатина — к.ф.-м.н., с.н.с., Институт математики СО РАН;  
e-mail: nataly@math.nsc.ru.*

*Дата поступления — 22 мая 2017 г.*