Полуавтоматическое пополнение словарей на основе синтаксических связей

Ю. В. Адаскина, П. В. Паничева, А. М. Попов

InfoQubes, InfoQubes, СПбГУ
ya@infoqubes.ru, pp@infoqubes.ru, ap@infoqubes.ru

Аннотация

Статья описывает разработанный нами метод полуавтоматического пополнения семантических классов на основе синтаксических связей в корпусе. Такой алгоритм необходим для снижения трудозатрат экспертов при разработке коммерческих приложений для автоматического анализа клиентских отзывов в интернете. Представленные результаты доказывают эффективность предложенного подхода.

1. Введение

Настоящая работа посвящена описанию метода полуавтоматического выделения слов, обозначающих персонал, из клиентских отзывов. Анализ мнений, опубликованных в интернете, является одной из наиболее приоритетных задач маркетинговых служб многих предприятий. При этом, для сервисных компаний зачастую центральным является вопрос анализа отзывов о персонале. Базой такой автоматической системы, основанной как на сентиментном анализе, так и на извлечении фактов, является словарь терминов персонала, обычно составляемый экспертом. Это довольно трудоемкая задача, так как требуется обработать большое количество документов, чтобы учесть все варианты обозначений персонала. Поскольку семантический класс терминов персонала в значительной степени зависит от предметной области, такую работу необходимо проводить для каждой новой тематики. Словари синонимов лишь несущественно облегчают подбор терминов и требуют дополнительного тестирования на релевантность.

Мы разработали метод полуавтоматического пополнения семантического класса терминов персонала, основанный на синтаксических связях в корпусе отзывов. При помощи такого алгоритма трудозатраты эксперта существенно снижаются. Данными для эксперимента послужил корпус текстов объемом 580 000 слов (около 45 000 уникальных лемм), состоящий из клиентских отзывов компании «S7 Airlines» и обработанный с помощью морфосинтаксического анализатора.

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

2. Предпосылки

Этот раздел посвящен обзору существующих подходов к решению похожих задач.

2.1. Алгоритмы пополнения семантических классов

Создание словарей для определенных семантических классов можно рассматривать как разновидность более общей задачи извлечения именованных сущностей. Извлечение именованных сущностей как таковое обычно включает выделение таких классов лексики, как наименования персон и организаций, географические названия и др. Изначально подходы к этой задаче опирались на лингвистические правила, создаваемые специалистом; современные системы в основном используют различные статистические методы и механизмы машинного обучения. Такие алгоритмы анализируют последовательные цепочки слов в предложении и приписывают наиболее вероятную разметку. Главными статистическими моделями, используемыми в таких системах, являются скрытые марковские модели и метод условных случайных полей; в других случаях задача распознавания именованных сущностей рассматривается как задача классификации по образцу, для чего используется, например, наивный байесовский классификатор или метод опорных векторов. Для использования всех этих алгоритмов необходим размеченный тренировочный корпус, и их эффективность во многом зависит от объема и качества разметки данных.

Кроме того, существует метод классификации с частичным привлечением учителя (bootstrapping), который позволяет изначально опираться на небольшой объем данных о семантическом классе и итеративно его пополнять. Примеры использования этого метода для извлечения сущностей описаны, например, в [2], [3], [5], [9].

В качестве параметров автоматического обучения алгоритмов пополнения словарей именованных сущностей могут использоваться синтаксические (например, [10], [12]) или семантические (например, [1], [11]) связи.

Важная особенность широко применяемых автоматических механизмов создания словарей именованных сущностей состоит в том, что обычно они

рассчитаны на очень небольшое количество основных сематических типов: персоны, организации, географические названия, названия валют. В работе [14] предлагается расширенная иерархия сущностей, включающая 200 типов (например, названия высших учебных заведений, фильмов, музыкальных групп и др.), на основе чего некоторые исследователи ([6], [4]) разработали автоматические методы создания детальных классификаторов.

2.2. Полуавтоматические итеративные алгоритмы пополнения семантических классов

Отдельно следует остановиться на работе [13], в которой решается задача полуавтоматического пополнения семантических классов. В исследовании описывается технология для автоматизированного создания лексиконов предметной области. В основу подхода положена гипотеза о том, что встречаемость слов исследуемой области в одних и тех же контекстах будет статистически значима. Подход состоит в том, чтобы имея небольшое число слов, точно относящихся к искомому семантическому классу, использовать эти слова для поиска других слов из этого класса.

На вход алгоритма подается корпус текстов из интересующей предметной области и небольшое множество слов, относящихся к интересующему семантическому классу. На выходе получается список слов, которые гипотетически принадлежат искомому семантическому классу, отсортированный по мере сходства.

Алгоритм состоит из следующих шагов:

- определить все предложения в тексте, где есть первоначальные релевантные слова; провести поверхностный синтаксический анализ (выделение базовых фразовых категорий);
- выделить окно просмотра вокруг первоначальных релевантных слов, которые являются вершинами фразовых категорий (окно [-1;1]), при этом выделяются только существительные, которые составляют так называемый категориальный контекст;
- для каждого слова из полученных категориальных контекстов вычисляется категориальный вес по формуле:
  $Score(W, C) = Fw_{context} / Fw_{corpus}$,
  где $Fw_{context}$ — это частота встречаемости слова в категориальных контекстах, а Fw_{corpus} — частота встречаемости слова во всём корпусе;
- удалить стоп-слова, числа и слова с общей частотностью не выше 5 (стоп-список включал 30 слов, в основном местоимения и определители), после чего полученный список сортируется по убыванию веса;
- пополнить список релевантных слов пятью самыми весомыми словами, которые не входят в первоначальный список релевантных, перейти к п.1.

Таких итераций проводилось несколько, до тех пор, пока не перестанут добавляться новые релевантные слова, или пока не выполнится заданное число итераций. Нужно отметить, что в работе рассматривалась возможность проведения эксперимента с учителем, когда на каждой итерации оценку вносимых новых релевантных слов выполнял бы человек, однако было решено исключить дополнительное вмешательство человека в процесс.

Эксперименты проводились на корпусе Fourth Message Understanding Conference (MUC-4), содержащий 1 700 текстов новостных статей по теме латиноамериканского терроризма (около полумиллиона слов), и было выбрано 11 семантических классов.

Алгоритм добавлял пять новых слов на каждой новой итерации. После завершения последней итерации строился окончательный список гипотетических слов, относящихся к целевому семантическому классу, упорядоченный по весу.

Список релевантных слов состоял из пяти для каждого класса для каждой итерации. Для оценки были вручную просмотрены 500 слов с максимальным весом из списка кандидатов в каждый семантический класс. В результате по теме латиноамериканского терроризма было отобрано 494 слова, или 9% от всеобщего объема просмотренных слов. Также приводится распределение плотности найденных слов в зависимости от числа просмотренных слов. Из этих данных видно, что для того, чтобы получить не менее 80% искомых слов от итогового списка, необходимо просмотреть примерно вдвое меньше слов, чем для получения 100%.

Представленный нами алгоритм относится к классу полуавтоматических итеративных алгоритмов пополнения семантических классов, но имеет несколько принципиальных отличий от вышеназванной работы.

Во-первых, в рамках нашей задачи по итеративному пополнению и дальнейшему использованию семантических классов необходимым условием является экспертная проверка всех выявленных слов для данного класса. Эта необходимость подтверждается также невысокой точностью работы таких алгоритмов. Поэтому мы приняли решение о подключении автоматической проверки полученных слов относительно экспертно созданного списка на этапе проведения итераций, а не только в ходе оценки результатов.

Во-вторых, мы рассматриваем задачу пополнения семантических классов как задачу классификации с информацией о позитивном классе. Поэтому мы используем сортировку не по весовым коэффициентам, а по коэффициентам близости, приписанным алгоритмом свободных векторов, опираясь на метод классификации на основе позитивного класса, представленный в [15].

3. Эксперимент

3.1. Первоначальные входные данные

Мы работали с неочищенным корпусом клиентских отзывов, объем которого составляет 580 000 слов, не считая знаки препинания (около 45 000 уникальных лемм). Для оценки полноты результатов используется разработанный экспертами словарь объемом в 100 слов. Еще один числовой показатель, необходимый для тестирования метода, — количество существительных и прилагательных в корпусе — 29 000. Естественно предполагать, что искомые термины персонала будут именами существительными, однако мы не должны исключать из анализа прилагательные, так как некоторые релевантные слова являются субстантивированными прилагательными (например, «служащий»), что не всегда корректно отражено в морфологической разметке. Основные количественные параметры представлены в следующей таблице:

Таблица 1. Данные

№	Тип данных	Значение
1	Общий размер корпуса (слова)	580 000
2	Количество уникальных слов в корпусе	45 000
3	Количество уникальных существительных и прилагательных в корпусе	29 000
4	Объем словаря, составленного экспертами	100

После исключения слов, встретившихся только один раз, а также не получивших при обработке синтаксических связей, осталось 23 100 существительных и прилагательных. Алгоритм пополнения семантического класса персонала основывается на поиске слов в корпусе, сходных по определенным параметрам с небольшим количеством первоначальных слов, обозначающих персонал и выбранных вручную. Так как значимой частью корпуса являются отзывы о работе персонала, мы предполагаем, что обозначения персонала окажутся среди самых частотных слов. Кроме того, выбор первоначальных слов из наиболее частотных даст больше входной информации для алгоритма пополнения, так как слова с высокой частотностью будут характеризоваться большим количеством существенных параметров.

Слово «бортпроводник» встретилось под номером 94 из наиболее частотных слов, слово «сотрудник» — под номером 103. Эти слова мы принимаем за первоначальные; их частотность в корпусе составляет, соответственно, 430 и 391.

3.2. Параметры классификации

В качестве параметров для алгоритма поиска сходных терминов используются синтаксические связи слов в корпусе. Наше исследование опирается на систему лингвистической обработки, которая производит морфосинтаксический анализ текста, основанный на правилах. В результате каждое предложение представляется в виде дерева зависимостей, где каждая синтаксическая связь принадлежит одному из 20 типов. Все имена существительные и прилагательные в корпусе представлены в виде распределения их синтаксических связей, где под связью понимается тройка (Тип связи, Роль данного слова, Второе слово в данной связи), в которой:

- Тип связи — название одного из 20 видов связей;
- Роль данного слова — бинарный параметр, обозначающий, является ли данное слово главным или зависимым в данной связи;
- Второе слово в данной связи — начальная форма слова на другой стороне связи, соответственно, зависимого, если данное слово является главным; или главного, если данное слово является зависимым.

По результатам предварительных экспериментов мы исключили из рассмотрения связи типа «Preposition» и «Undefined», соответственно, связи с предлогами и с союзами в сочинительных группах. Также были исключены все связи, которые встретились в корпусе только один раз. Для каждой связи данного слова ее вес нормализуется относительно количества всех связей данного слова. Для первоначальных входных слов мы получили, к примеру, следующие параметры:

Таблица 2. Параметры для слова «Сотрудник»

№	Тип связи	Роль в связи	Связанное слово	Вес
1	DirectObject	Target	БЛАГОДАРИТЬ	0.003
2	IndirectObject	Target	БЛАГОДАРНОСТЬ	0.004
3	Posessive	Source	ВАШ	0.17

Таблица 3. Параметры для слова «Бортпроводник»

№	Тип связи	Роль в связи	Связанное слово	Вес
1	Posessive	Source	ВАШ	0.001
2	Subject	Target	ОТКАЗАТЬ	0.008

3.3. Исключение «сильно негативных» слов

На самом первом этапе классификации из списка имен существительных и прилагательных были автоматически выбраны слова, не содержащие ни одной связи из тех, которые встретились с двумя первоначальными («бортпроводник» и «сотрудник»).

Мы следовали принципу выявления «сильно негативных слов», описанному в [15] как выявление «сильно негативных документов». «Сильно негативных слов» оказалось 17 600, оставшихся, полезных для эксперимента слов — 5 500.

3.4. «Проверка» слов на каждом шаге

На каждой итерации алгоритма модель классификации корректируется в соответствии с полученными релевантными словами (подробнее см. 3.6), отобранными на основании экспертного списка из 100 слов. Мы ожидаем, что такая корректировка позволит улучшить результат по отношению к базовой классификации, основанной на одной первоначальной модели.

3.5. Базовый алгоритм

В качестве базового алгоритма для сравнения результатов мы произвели одну классификацию без итераций. В обучающую выборку вошли «сильно негативные слова», а также 103 наиболее частотных слова из первоначальных входных данных, из которых 2 были размечены как «позитивные» (а именно, «бортпроводник» и «сотрудник»), а оставшееся 101 слово считалось «негативным». В качестве тестовой выборки были использованы 5 400 слов, не вошедших ни в «сильно негативные», ни в 103 слова, обработанных вручную. Для статистической обработки мы применили алгоритм опорных векторов ([8]) с линейным ядром.

Результаты были отсортированы по полученному параметру близости к «позитивному» классу. Так как базовый алгоритм необходим для сравнения с результатами итеративного метода, отсортированный результат работы базового алгоритма был разбит на группы по 50 слов; в качестве базового результата мы оценивали количество релевантных слов, которое встречается в каждой группе.

3.6. Итеративный полуавтоматический алгоритм пополнения семантического класса

Алгоритм полуавтоматического пополнения основан на предположении о том, что обработка результатов на каждой итерации и включение в новый «позитивный» класс только релевантных слов позволит получить большее количество релевантных результатов при проверке меньшего объема слов.

Мы экспериментировали с различными объемами проверяемых слов на каждом шаге N={10, 25, 50, 100, 150}. Цикл работы алгоритма описывается следующим образом:

1. На основе обучающих данных запускается тренировка алгоритма опорных векторов.

2. Затем происходит тестовая итерация модели на оставшихся неизвестных словах. Из полученного результата автоматически выбирается N слов, наиболее близких к классу положительных.

3. На каждом шаге в группе N слов выбираются релевантные в соответствии с экспертно составленным списком, остальные считаются отрицательными. Обучающая выборка пополняется соответственно новыми положительными и отрицательными словами. Процесс повторяется заново, начиная с п.1.

В рамках предварительного эксперимента мы опирались на работу [13], с отличиями в способе классификации (метод опорных векторов вместо сортировки по весу) и набором параметров (синтаксические связи вместо соседних имен существительных). Однако результаты этого эксперимента на наших данных неудовлетворительны, как мы полагаем, из-за особенностей нашего корпуса данных: во-первых, он содержит неочищенные пользовательские отзывы без предварительной обработки и исправления ошибок и опечаток; во-вторых, работа персонала не является тематикой всех документов, поэтому объем полезной части корпуса существенно снижается. Поэтому мы предлагаем дальнейшее развитие идеи, более подходящее, на наш взгляд, для обработки русского языка и небольшого неочищенного корпуса данных: мы считаем, что если оценивать кандидаты в релевантные слова относительно экспертного списка на каждой итерации, то можно добиться более высокого качества при меньших трудозатратах. Таким образом, мы выделяем четыре направления для развития оригинального алгоритма:

- использовать полный синтаксический анализ вместо частичного;
- вместо единой частотной метрики [13] использовать метод опорных векторов по параметрам, характеризующимся типом связи, направлением связи, леммой;
- задействовать проверку гипотетических релевантных слов на каждой итерации;
- на каждом шаге пополнять не только положительный, но и отрицательный класс слов.

4. Результаты

Нас интересуют результаты работы алгоритма при объемах слов, проверенных относительно экспертного класса, сильно меньших, чем количество потенциально релевантных слов, то есть всех слов со связями, исключая «сильно негативные». Так как количество потенциально релевантных слов около 5 000, мы исследовали результаты работы итеративного алгоритма, а также базового алгоритма, до 2500 слов. Для каждого значения итерации была получена следующая таблица:

Таблица 4. Полученные данные для эксперимента с шагом в 150 слов

Номер итерации	Количество релевантных слов	Количество проверенных слов	Суммарная релевантность итерации	Суммарное количество релевантных слов	Суммарное количество проверенных слов	Средняя релевантность итерации по всем шагам
1	15	150	0.1	15	150	0.1
2	11	150	0.0733	26	300	0.0867
3	14	150	0.0933	40	450	0.0889
4	18	150	0.12	58	600	0.0967
5	4	150	0.0267	62	750	0.0827
6	10	150	0.0667	72	900	0.08
7	7	150	0.0467	79	1050	0.0752
8	1	150	0.0067	80	1200	0.0667
9	4	150	0.0267	84	1350	0.0622
10	1	150	0.0067	85	1500	0.0567
11	1	150	0.0067	86	1650	0.0521
12	0	150	0.0	86	1800	0.0478
13	1	150	0.0067	87	1950	0.0446
14	0	150	0.0	87	2100	0.0414
15	1	150	0.0067	88	2250	0.0391
16	0	150	0.0	88	2400	0.0367
17	0	150	0.0	88	2550	0.0345

Для итеративного алгоритма для всех примененных объемов итерации N, а также для базового неитеративного алгоритма, были проанализированы[1] следующие параметры:

- сколько релевантных слов дополнил алгоритм при увеличении объема всех проверенных слов (Рисунок 1);
- как меняется средняя релевантность алгоритма, выраженная через отношение всех полученных релевантных ко всем проверенным, при увеличении объема проверенных слов (Рисунок 2).

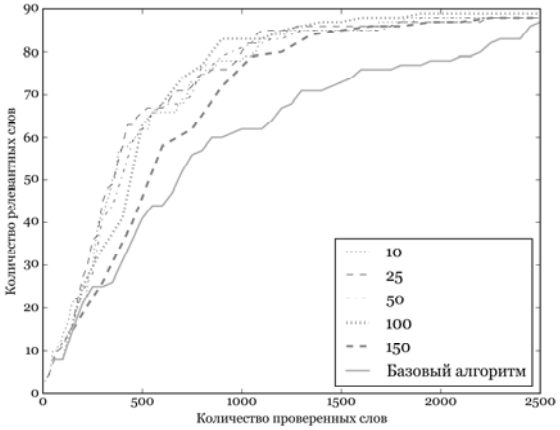

Рис. 1. Зависимость количества релевантных слов от количества просмотренных слов для различных значений шагов

Из графика на Рисунке 1 видно, что достаточно высокая полнота (более 80%) достигается при некоторых итерациях на объеме проверенных слов 1 000. На Рисунке 2 такому объему соответствует точность алгоритмов около 0.1. Средняя точность алгоритма будет расти при уменьшении объема проверенных слов и понижении полноты и достигнет около 0.15 при объеме проверенных слов 500 и полноте около 65%.

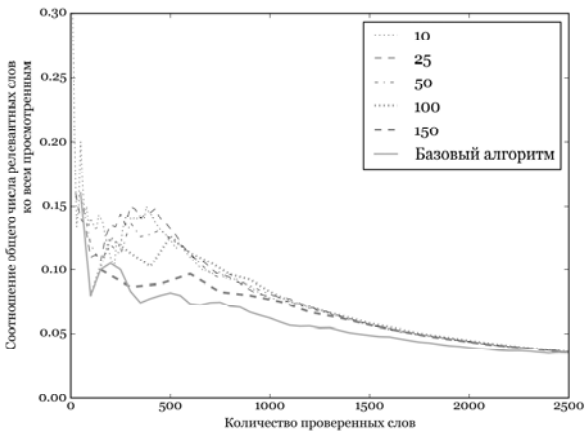

Рис. 2. Зависимость средней релевантности алгоритма (соотношения релевантных слов ко всем просмотренным) от количества просмотренных слов для различных значений шагов

И числовые значения, и общие тенденции сравнимы с результатами, описанными в [13].

5. Выводы

Полученные результаты показывают применимость предлагаемого алгоритма к задаче полуавтоматического пополнения семантических классов. Алгоритм позволяет достичь приемлемого качества при анализе небольшого специализированного корпуса, состоящего из предварительно не очищенных данных, различных по их отношению к тематике рассматриваемого семантического класса. Алгоритм позволил выявить редкие, неочевидные варианты

[1] Была использована программная библиотека [7].

обозначения персонала и слова с опечатками, к примеру, такие как «девушка на стойке регистрации», «IT-шник» и «опертор», которые было бы крайне сложно определить экспертно.

Результаты при небольшом значении шага превосходят результаты с крупным шагом при невысоких объемах всех просмотренных слов. При увеличении суммарного объема просмотренных слов результаты алгоритмов с различным шагом сближаются. Таким образом, выбор оптимального алгоритма зависит от желаемой полноты результата и имеющегося времени на разметку слов (релевантность ~0.15 при просмотре 500 слов с шагом 25, ~0.085 при просмотре 1 000 слов с шагом 100). Кроме того, в таком эксперименте при любом значении шага итеративный алгоритм выдает результаты лучше, чем базовый, который не повторяет классификацию с учетом новой разметки. Достаточно высокая полнота (90 слов из 100) достигается при любом шаге при просмотре 2 500 слов.

Для повышения средней релевантности алгоритма предлагается в дальнейшем включить новые параметры для классификации, такие как морфологические, орфографические и контекстные; исследовать роль фильтрации параметров; а также применить другие статистические методы.

Литература

[1] Кузнецов И. Методики выявления объектов и связей, заданных в неявном виде. Труды международной конференции “Диалог 2013”.

[2] Abney S. Bootstrapping // 40th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference, 2002.

[3] Becker M., Hackey B., Alex B., Grover C. Optimising Selective Sampling for Bootstrapping Named Entity Recognition // Proceedings of the Workshop on Learning with Multiple Views, 22nd ICML, Bonn, 2005.

[4] Chang J., Tsai R. T.-H., Chang J. S. Wikisense: Supersense tagging of Wikipedia named entities based WordNet // Proceedings of PACLIC, 2009.

[5] Collins M., Singer Y. Unsupervised models for named entity classification // Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[6] Higashinaka R., Sadamitsu K., Saito K., Makino T., Matsuo Y. Creating an Extended Named Entity Dictionary from Wikipedia // Proceedings of COLING 2012: Technical Papers, Mumbai, December 2012.

[7] Hunter J. D. Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering, 9 (2007), Nr. 3.

[8] Joachims T. Making large-Scale SVM Learning Practical // Advances in Kernel Methods — Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999.

[9] Kozareva Z. Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists // EACL, The Association for Computer Linguistics, 2006.

[10] Mavljutov R. R., Ostapuk N. A. Using basic syntactic relations for sentiment analysis // Proceedings of The International Conference “Dialog”, 2012.

[11] McKeown K., Hatzivassiloglou V. Augmenting Lexicons Automatically: Clustering Semantically Related Adjectives // Proceedings of ARPA Workshop on Human Language Technology, 1998.

[12] Mohit B., Hwa R. Syntax-based Semi-Supervised Named Entity Tagging // ACL, 2005.

[13] Riloff E., Shepherd J. A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction // Natural Language Engineering 5(2), 1999.

[14] Sekine S., Sudo K., Nobata C. Extended named entity hierarchy // Proceedings of LREC, 2002.

[15] Yu H., Han J., Chang K.C.C. PEBL: Positive Example-Based learning for web page classification using SVM // Proceedings of ACM SIGKDD, 2002.

Semi-Automatic Lexicon Augmenting Based on Syntactic Relations

Yu. V. Adaskina, P. V. Panicheva, A. M. Popov

The paper presents a semi-automatic algorithm for domain-specific semantic classes compilation. We are dealing with raw user reviews analysis for service companies, for whom staff evaluations are a major concern. The first step of any analytical system dealing with such data, intended to carry out either opinion mining or fact extraction, is creating a staff terms lexicon. The aim of our research was to reduce expert work while creating staff terms lexicon for commercial applications. We use a syntactically annotated corpus of user reviews on airline service. The method is based on two staff terms that have come up in 103 most frequent words list. They constitute the initial positive class, while the other 101 are regarded as the initial negative class. Using syntactic relations as machine learning parameters we bootstrap new words and build up the lexicon. The results show that the algorithm is indeed helpful for reducing expert’s time and effort and allows for up to 90% recall. Experiment parameters can be adjusted depending on how many staff terms need to be covered and how much time is available to the expert.