

БАЙЕСОВСКИЙ КЛАССИФИКАТОР ИМЕНОВАННЫХ СУЩНОСТЕЙ В ТЕКСТЕ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Представлен метод распознавания и классификации именованных сущностей. В методе используется подход обучения с учителем. Построен Байесовский классификатор, позволяющий определять класс сущностей, представленных с помощью векторов бинарных значений признаков. В отличие от наивного Байесовского подхода, осуществляется отход от предположения о независимости признаков за счет использования в работе классификатора полной базы примеров из учебной коллекции.

Понятие именованной сущности (*named entity*) было сформулировано в 1990-х гг. на конференции по пониманию сообщений (Message Understanding Conferences), посвященной проблемам извлечения информации из текстов. Под именованными сущностями понимаются элементы текста, обозначающие персону, название организации, обозначение места, а также некоторые числовые данные: даты, время, валюту, процентные данные [1].

Таким образом, понятие именованной сущности пересекается с ономастикой (так как первичная номинация персон, организаций и мест выполняется с помощью имен собственных). Синтаксически именованные сущности, как правило, выражены именными группами, включающими именованную сущность, представленную главным или определяющим существительным. В редких случаях — отсылка на объект именованной сущности осуществляется путем деноминативных прилагательных (европейский, викторианский и т. п.).

Задача идентификации именованных сущностей (при рассмотрении «текстовых», в отличие от «числовых», сущностей, таких как персона, место, организация) для большинства европейских языков решается как поиск в тексте лексических единиц с капитализацией первого символа. В дальнейшем для решения задачи извлечения фактов эту процедуру можно усложнить, дополнив поиском кореферентных лексических единиц.

Далее встает задача классификации идентифицированных объектов — отнесение к одному из классов именованных сущностей.

Формально задачу классификации можно представить следующим образом. Рассматривается текст на естественном языке. Отдельные слова и устойчивые словосочетания далее будем называть объектами. Пусть O — множество всех объектов, C — множество классов именованной сущности, к которому может быть отнесен объект $o \in O$ в анализируемом тексте.

При решении задачи распознавания именованной сущности (объекта) $o \in O$ в корпусе текстов определяется набор признаков $\bar{f} = \{f_t\}_{t \in T}$ на основании которых объекты $o \in O$ будут сопоставляться. В общем случае, признак f_t может иметь булево, числовое или

номинативное значение. Мы рассматриваем только булевы признаки:

$$(\forall o \in O) \left(f_t(o) = \begin{cases} 1, & \text{если признак } t \text{ выполняется для } o, \\ 0, & \text{в противном случае.} \end{cases} \right).$$

Объекты o_1, o_2 : $\bar{f}(o_1) = \bar{f}(o_2)$ считаются эквивалентными. При невозможности отождествления объектов o_1, o_2 : $\bar{f}(o_1) = \bar{f}(o_2)$ необходимо расширение множества T используемых признаков. Различным классам $c_1, c_2 \in C$ соответствуют непересекающиеся наборы признаков:

$$c_1 \neq c_2 \Rightarrow \{\bar{f}(o): o \in c_1\} \cap \{\bar{f}(o): o \in c_2\} = \emptyset.$$

Возникновение коллизий показывает необходимость расширения множества классов.

Таким образом, задача распознавания именованных сущностей представляет собой задачу классификации объектов.

Для решения этой задачи существует следующие подходы.

1. Рационалистический подход: идентификация именованных сущностей осуществляется на основе продукционных правил, обычно собранных вручную. Правила, сформированные на основании лингвистического анализа корпуса текстов, позволяют создать шаблоны (например, в формате регулярных выражений), которые дают возможность идентифицировать именованные сущности на основе лексико-грамматических признаков текстовых единиц, их синтактики, орфографии, а также составить словники, необходимые для работы правил идентификации (см., например, [2–4]). Разрабатываются ряд проектов (например, Ontos-Miner, ИСИДА-Т) для русского и некоторых европейских языков, которые позволяют на основе применения синтаксических анализаторов, а также семантических словарей и онтологий, выделять из текстов интересующие пользователей сущности и идентифицировать отношения между ними.

2. Машинное обучение: задача поиска правил для идентификации именованных сущностей формулируется как решение задачи классификации с использованием статистических моделей. Методы машинного обучения, используемые для решения задачи, делятся на несколько этапов.

2.1. Обучение «с учителем»: обучение на основе учебной коллекции, включающей явно специфици-

рованные (вручную) именованные сущности. Методы этой группы оценивают параметры для положительно определенных примеров корпуса и при работе с новым корпусом используют значения этих параметров. Сюда относятся Байесовский классификатор, скрытые марковские модели, принцип максимума энтропии, деревья принятия решений, метод опорных векторов, условные случайные поля и др.

2.2. Частичное обучение «с учителем»: от предыдущего подхода отличается тем, что исходная учебная коллекция содержит очень маленький набор начальных данных. При помощи реализации метода бутстреппинга осуществляется итеративное обучение классификатора.

2.3. Обучение «без учителя»: для решения задачи не требуют предварительного создания корпуса примеров. Такие методы способны сделать вывод по «сырому» текстовому материалу.

3. Гибридный подход: объединяет методы двух подходов. Для выявления набора признаков применяется лингвистический анализ, далее числовые показатели, соотносящиеся с выявленными признаками, определяются с помощью методов машинного обучения.

В данной работе для решения задачи классификации определена функция Classifier: $O \rightarrow C$, которая находит класс $c \in C$, которому соответствует максимальная степень принадлежности объекта $o \in O$.

Ключевым компонентом здесь является этап определения множества признаков, оказывающих влияние на классификацию объектов — это основная задача, которая решается индивидуально, и, в конечном счете, выбор признаков, при сходстве подходов к оценке, предопределяет качество работы конкретного алгоритма. Для задачи распознавания и классификации именованных сущностей релевантными признаками могут являться капитализация, наличие в составе слова специальных символов, частота, наличие определенных контекстов, позиция в предложении, наличие определенных морфологических элементов в составе слова, грамматические характеристики слова и т. д. (в [5] представлен довольно полный обзор часто используемых признаков).

Для построения классификатора составляется учебная коллекция $\tilde{O}$ текстов, где в явной форме для некоторых объектов $o \in \tilde{O}$ обозначено их соответствие определенному классу $c \in C$.

Индикатором принадлежности объекта $n \in O$, эквивалентного некоторому объекту $o \in \tilde{O}$, заданному классу $c \in C$ является функция

$$\begin{aligned} \text{BoolClassifier}(n, c) &= \bigcup_{o \in c} (\bar{f}(n) = \bar{f}(o)) = \\ &= 1 - \prod_{o \in c} \max_{t \in T(o)} |f_t(n) - f_t(o)|, \end{aligned}$$

где $T(o) \subset T$ — множество значимых признаков для объекта $o \in c \cap \tilde{O}$.

Легко проверить, что из существования такого $o \in c \cap \tilde{O}$, что $\bar{f}(n) = \bar{f}(o)$ следует $\text{BoolClassifier}(n, c) = 1$,

а в противном случае $\text{BoolClassifier}(n, c) = 0$. Функция BoolClassifier не позволяет определить класс объектов $o \in O$ вне имеющих эквивалентных представителей в учебной коллекции $\tilde{O}$. Следовательно, данный индикатор не может быть использован для построения функции Classifier.

Рассмотрим нечеткий индикатор

$$\text{FuzzyClassifier}(n, c) = 1 - \prod_{o \in c} \left[\frac{1}{|T(o)|} \sum_{t \in T(o)} |f_t(n) - f_t(o)| \right].$$

Утверждение

$(\forall n \in O, \forall c \in C) \times$

$\times (\text{BoolClassifier}(n, c) \leq \text{FuzzyClassifier}(n, c) \leq 1)$.

Доказательство. Первое неравенство есть следствие того, что максимальное значение элементов числового множества не меньше их среднего арифметического. Второе неравенство очевидно. Утверждение доказано ■

Легко убедиться, что при полном совпадении n с одним из $o \in c \cap \tilde{O}$ значение индикатора FuzzyClassifier совпадает со значением BoolClassifier и равно 1. Каждый сомножитель произведения в формуле для FuzzyClassifier представляет собой долю несовпадающих координат информационных векторов $\bar{f}(n)$ и $\bar{f}(o)$, $o \in c \cap \tilde{O}$, то есть может быть интерпретирован как мера различия объектов n и o или как вероятность $\mathbb{P}\{n \neq o : o \in c \cap \tilde{O}\}$. Тогда все произведение можно интерпретировать как степень того, что объект n не эквивалентен ни одному из объектов $o \in c \cap \tilde{O}$, или вероятность $\mathbb{P}\{\bar{f}(n) \mid n \notin c\}$ того, что объект n , не принадлежащий классу c , характеризуется определенным вектором признаков $\bar{f}(n)$. Следовательно, индикатор $\text{FuzzyClassifier}(n, c)$ интерпретируется как вероятность противоположного события, состоящего в том, что вектор признаков характеризует объект n , который принадлежит классу c , или как условную вероятность $\mathbb{P}\{\bar{f}(n) \mid n \in c\}$, или проще $\mathbb{P}\{\bar{f} \mid c\}$.

В терминах условных вероятностей задача классификации может быть переформулирована как поиск вероятности $\mathbb{P}\{c \mid \bar{f}\}$ того, что класс c является истинным классом для объекта n , при условии что последний представлен вектором признаков $\bar{f}(n)$. Таким образом, нечеткий индикатор $\text{FuzzyClassifier}(n, c)$ может быть использован для построения классификатора на основе теоремы Байеса:

$$\text{Classifier}(n) = \arg \max_{c \in C} [\text{FuzzyClassifier}(n, c) \cdot |c \cap \tilde{O}|].$$

Классификатор FuzzyClassifier использует в своей работе примеры объектов класса c , что позволяет учитывать вероятности не отдельных признаков, а их наборов для репрезентации объектов из класса c .

Для реализации функции $\text{FuzzyClassifier}(n, c)$ Classifier: $O \rightarrow C$ разрабатывается комплекс программ для ЭВМ. В состав комплекса входят:

– библиотека T тестеров: каждый тестер представляет функцию $t(o)$, определяющую истинность некоторого утверждения о характеристиках переданного объекта $o \in O$;

– база данных $\tilde{O}$ объектов учебной коллекции: для $o \in \tilde{O}$ каждого объекта определены его класс, список значимых тестов и их значений;

– функция $\text{FuzzyClassifier}(o)$, возвращающую класс, степень принадлежности которому объекта o максимальна; данная функция также собирает статистику о классах.

Библиографический список

1. Sundheim, Beth M. Overview of Results of MUC-6 Evaluation [Электронный ресурс] / Beth M. Sundheim // Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference (Columbia, Maryland, November 6–8 1995). Morgan Kaufman Publishers Inc., 1995. P. 13–32. URL: <http://aclweb.org/anthology/M/M95/M95-1002.pdf>.

2. Lappin, S. An Algorithm for Pronominal Anaphora Resolution / Shalom Lappin, J. Leass Herbert // Computational Linguistics. 1994. Vol. 20, №. 4. P. 535–561.

3. Bontcheva, K. Cunningham. Shallow Methods for Named Entity Coreference Resolution [Электронный ресурс] / Kalina Bontcheva, Marin Dimitrov, Diana Maynard at el. // Chaînes de references et solveurs d'anaphores, workshop TALN (Nancy, 24–27 June 2002). URL: <http://user.phil-fak.uni-duesseldorf.de/~rumpf/SS2008/IE/Pub/BonDimMay02.pdf>.

4. Sparks, N. L. UDel: Named Entity Recognition and Reference Regeneration from Surface Text / Nicole L. Sparks, Charles F. Greenbacker, F. Kathleen at el. // Proceedings of the 6th International Natural Language Generation Conference (Trim, Co. Meath, Ireland, July 7–9 2010). Dublin, 2010. P. 241–242.

5. Nadeau, D. A. Survey of Named Entity Recognition and Classification / D. A. Nadeau, Satoshi Sekine // Recognition, Classification and Use, Benjamins Current Topics 19. NY ; Lisbon : John Benjamins Publishing, 2009. P. 3–27.