ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ т. 9, №4(39), с. 3–33

ББК 32.813
ГРНТИ 16.31.21, 20.19.27
УДК 004.89:004.912

И. В. Трофимов, Е. А. Сулейманова, Н. А. Власова,
А. В. Подобряев

Разрешение событийно-несобытийной неоднозначности существительных

Аннотация. В статье рассматривается проблема автоматического различения событийных и несобытийных употреблений многозначных существительных, актуальная в контексте задачи извлечения событийной информации из текста. Предложен метод разрешения событийно-несобытийной неоднозначности, опирающийся на автоматически построенное обучающее множество. Для обучения классификатора используется множество примеров, содержащих однозначно событийные и однозначно несобытийные существительные, дистрибутивно близкие целевому неоднозначному существительному. Метод дистрибутивных замен позволил получить удовлетворительные результаты на небольшой выборке неоднозначных существительных. Результаты подтвердили значимость синтаксического контекста при решении данной задачи. Дальнейшее развитие метода предполагает построение более совершенной дистрибутивной модели и полную автоматизацию процедуры подбора замен.

Ключевые слова и фразы: разрешение лексической неоднозначности, автоматическое построение обучающего множества, модель дистрибутивной семантики, событие, событийное существительное, извлечение событийной информации.

Введение

При извлечении из текста информации о событиях одной из подзадач является обнаружение маркеров целевых событий — слов, посредством которых упоминаются ситуации, представляющие интерес с точки зрения задачи извлечения. (В скобках заметим, что термин *ситуация* используется в лингвистике в качестве общего обозначения разных непредметных сущностей — состояний, явлений, действий, свойств. В таком же обобщенном значении может употребляться

Работа выполнена по НИР «Исследование и разработка методов автоматического извлечения событийно-темпоральной информации из текстов» в рамках гос. задания, номер гос. регистрации АААА-А17-117040610371-7.

© И. В. Трофимов, Е. А. Сулейманова, Н. А. Власова, А. В. Подобряев, 2018
© Институт программных систем имени А. К. Айламазяна РАН, 2018
© Программные системы: теория и приложения (дизайн), 2018

10.25209/2079-3316-2018-9-4-3-33

и термин *событие*. Именно в таком — метаязыковом — смысле используется термин *событие* в настоящей работе.)

Выявление маркеров событий в тексте предполагает решение целого ряда проблем: очертить в терминах тематических и/или акциональных категорий круг предикатов, которые в рамках общей задачи будут относиться к целевым событиям; определить ограничения на референциальный и фактуальный (отражающий отношение к действительности) статус целевой пропозиции; построить и реализовать алгоритмы для выявления соответствующей информации в тексте.

В настоящей статье рассматривается один специфический аспект упомянутой задачи.

Для референции к событиям в тексте служат не только глаголы в различных формах, но и т. наз. «событийные существительные»: *война*, *собрание*, *выборы*, *поездка*, *переговоры*, *конфликт*. **Событийное существительное** — существительное, обозначающее событие (ситуацию). Альтернативные названия этого класса лексики — *событийные имена* [**1**], *предикатные имена* [**2**] (у этого термина существует и другое употребление — в значении «существительное, образованное от глагола и прилагательного» [**3**]), *событийные номинализации* [**4**], *nomina actionis*, *event nominals*. Событийные существительные образуют класс лексики, пересекающийся, но не совпадающий с классом отглагольных (отпредикатных) имен, или девербативов. С одной стороны, есть событийные существительные, которые не относятся к отглагольным: это имена ситуаций, не являющиеся отглагольными дериватами (*свадьба, вахта, спектакль*) или образованные от глагольной основы, но не по стандартной модели (*война, любовь, сделка*). С другой стороны, далеко не все отглагольные существительные являются событийными: от глагольной основы образуются не только имена ситуаций, но и имена участников ситуаций (*слушатель, будильник, ловушка*).

Автоматическое выявление маркеров событий значительно осложняется тем обстоятельством, что существительным, способным выступать в роли имен ситуаций (событий, процессов, действий, состояний), свойственна регулярная многозначность [**5**] — это в основном метонимические переносы от имени ситуации к именам ее участников. Наиболее типичные несобытийные значения событийных существительных приведены в таблице 1.

При выявлении событий в тексте возникает необходимость отличить существительное в событийном значении от его несобытийных употреблений.

Таблица 1. Несобытийные значения, свойственные неоднозначным событийным существительным

Несобытийные значения	Примеры
субъект	*защита, охрана, управление*
результат (и субъект)	*отложение, скопление, уплотнение*
результат (и объект)	*выпечка, выписка, добыча, завтрак, заготовка, заимствование, издание, нарезка, обед, организация, открытие, отправление, подвеска, поджарка, покупка, постройка, посылка, публикация, сооружение, сочинение*
результат (и место; но не объект)	*вывих, повреждение, порез, разрушение, разрыв, ушиб*
результат (и способ; но не объект)	*анализ, аранжировка, изложение, изображение, интерпретация, исполнение, исследование, критика, описание, организация, осмысление, перевод, подача, покраска, понимание, постановка, редакция, решение, стрижка, толкование, трактовка, употребление*
средство или «второй объект» [5]	*вооружение, документация, замазка, заправка (кулинарн.),* начинка, обертка, обивка, обшивка, побелка, подкормка, упаковка, шпатлевка, штукатурка
инструмент	*зажигание, звонок, освещение, отопление*
место	*выезд, вырубка, заправка, обрыв, стройка, стоянка, химчистка*

1. Подходы к автоматическому разрешению многозначности

В области компьютерной обработки текста термином *word sense disambiguation (WSD)* обозначается автоматическое распознавание значения (многозначного) слова в контексте его употребления. Аналогичного общепринятого русского термина, по-видимому, пока не существует — ср. варианты *автоматическое разрешение лексической многозначности*, *снятие лексической неоднозначности*.

1.1. WSD: варианты постановки задачи, методы

С середины XX века проблема различения значений слов (the problem of multiple meaning, ambiguity) считается серьезнейшим препятствием на пути компьютерной обработки естественного языка. Впервые она упоминается как фундаментальная проблема машинного перевода

уже в 1949 году в меморандуме Уоррена Уивера [6], который положил начало исследованиям в области использования компьютеров для перевода с языка на язык. Уиверу принадлежит и идея использования контекстного окна для разрешения неоднозначности. В 1950 году Абрахам Каплан из Rand Corporation исследовал микроконтексты многозначных слов в математических текстах с целью установить их (контекстов) оптимальный размер. Полученные результаты были скромны, но вселяли надежду на принципиальную разрешимость проблемы многозначности слов при машинном переводе, а также свидетельствовали о пользе статистических данных для этой цели [7].

Известные на сегодняшний день методы автоматического разрешения лексической неоднозначности различаются по двум группам параметров. Параметры первой группы связаны с постановкой задачи.

- Охват:
  - выборочный метод (*lexical sample WSD, targeted WSD*);
  - универсальный метод (*all-words WSD*).
- Наличие или отсутствие заранее заданного перечня (инвентаря) значений слов:
  - когда инвентарь значений изначально определен, разрешение лексической неоднозначности представляется как задача классификации, а цель состоит в том, чтобы установить, к какому из значений относится слово в конкретном употреблении, т. е. контексте (такая постановка задачи именуется *word sense disambiguation, WSD*);
  - если инвентарь значений изначально не известен, то целью ставится кластеризация употреблений слова; при этом контексты употребления слова группируются в кластеры, каждому из которых, как предполагается, соответствует некоторое значение слова (такой подход называют *word sense induction (WSI)*, реже *word sense discrimination*).
- Для задачи WSD — степень детализации значений (granularity) в инвентаре (подробнее об этом в отдельном разделе).

Вторая группа параметров характеризует собственно методы разрешения неоднозначности.

- Выбор способа представления контекста. Контекст принято представлять в виде вектора признаков, среди которых выделяют следующие:
  - локальные и тематические. Данное противопоставление различает признаки, извлекаемые из ближайшей окрестности (окна) многозначного слова, и признаки, извлекаемые из текста

в целом, либо какого-то его крупного фрагмента (параграфа). Предполагается, что последние характеризуют тему текста (фрагмента) в целом;
  – поверхностные, синтаксические и семантические. Разделение на эти группы мотивировано традиционным выделением в языке нескольких уровней. К поверхностным признакам можно отнести лексический состав контекста, информацию о частях речи, информацию о расположении контекстного слова относительно многозначного слова. К синтаксическим — лексический состав контекста с учетом типа синтаксического отношения, связывающего контекстное слово с многозначным. К числу семантических относят такие признаки, которые были выявлены с использованием информационных ресурсов, содержащих сведения о значениях, толкованиях и семантических отношениях между словами.

- Выбор метода разрешения неоднозначности. Для решения WSD-задачи методологическое ядро формируют две крупные группы:
  – методы, разрешающие неоднозначность путем сопоставления контекста многозначного слова с содержимым какого-либо внешнего лексикографического или лексико-семантического ресурса (например, с толкованием в толковом словаре или с парадигматической окрестностью многозначного слова в тезаурусе);
  – методы, разрешающие неоднозначность посредством классификатора, обученного на (так или иначе) аннотированном корпусе текстов. Применяются практически все современные обучаемые методы классификации: метод опорных векторов; наивный байесовский классификатор (НБК); деревья и списки решений; k-ближайших соседей; нейросетевые классификаторы, а также различные модели комбинирования классификаторов (т. наз. ensemble methods).
- Используемые алгоритмами ресурсы:
  – структурированные ресурсы знаний — словари, онтологии, тезаурусы;
  – неструктурированные ресурсы — размеченные и неразмеченные корпуса текстов.

Подробнее о состоянии дел в области WSD см. обзоры Роберто Навильи [**8**, **9**].

1.2. Методы, применяемые для автоматизации построения обучающего множества

Пальму первенства в WSD удерживают методы обучения с учителем (supervised learning). Единственным узким местом этих методов является их потребность в огромных объемах текста, в которых для всех слов или группы слов выполнена разметка значений. Получение таких корпусов (sense-tagged corpora) связано с большими затратами квалифицированного труда — эта проблема известна под названием knowledge acquisition bottleneck. Попытки по возможности снизить трудозатраты на аннотирование текстов, превратив обучение с учителем (supervised learning) в частичное обучение с учителем (semisupervised learning), предпринимались в нескольких направлениях.

Технология **бутстрэппинга** позволяет получить необходимые для обучения объемы данных с помощью небольшого начального множества, построенного вручную. В области WSD применяются два вида бутстрэппинга: самообучение и совместное обучение [**10**, **11**].

Активное обучение снижает затраты ручного труда за счет автоматического отбора наиболее информативных примеров для последующего ручного аннотирования [**12**, **13**].

Для получения обучающих данных практически без затрат на ручную разметку применяются методы, основанные на использовании разного рода **замен**: значение многозначного слова подменяется чем-то, что может быть размечено в большом объеме текста без особых усилий.

Так, в качестве обучающих данных могут использоваться большие многоязычные массивы параллельных текстов (за значение многозначного слова принимается его перевод на другой язык) [**14**]. В другой работе [**15**] описано построение обучающего массива с помощью тезауруса Роже и энциклопедии Гролье. Отдельное значение многозначного слова моделируется как соответствующая ему категория в тезаурусе, а обучающее множество формируется из предложений энциклопедии, содержащих другие слова этой категории. Метод тестировался на 12 существительных и показал очень хорошие результаты (точность 95–99%) на омонимах типа *crane* — ‘журавль’ и ‘кран’ (тестовое множество формировалось также на основе энциклопедии Гролье). Однако другими исследователями высказывались сомнения относительно того, насколько хорошо такой классификатор, обученный на строго тематически организованном материале, будет работать на обычном тексте.

Для получения релевантных контекстов для разных значений многозначного слова могут использоваться однозначные лексически

близкие им слова и словосочетания (метод т. наз. **monosemous relatives**). Источником таких «родственников» обычно служит WordNet [**16**]. Например, у одного из значений полисемичного слова *suit* есть однозначный гипоним *buisiness suite* ‘деловой костюм’, а у другого — однозначный гипероним *legal proceeding* ‘судопроизводство’.

Приводятся данные [**17**], что в корпусе, содержащем 8,5 тыс. многозначных слов с суммарным числом значений 24 тыс., у 64% значений имеются в WordNet однозначные синонимы, непосредственные гипонимы или гиперонимы. Предполагается, что в корпусе большего размера доля таких случаев выше; кроме того, ее можно увеличить за счет подключения других лексических отношений (например, меронимии, антонимии и т. п.).

В эксперименте, поставленном Ликок и др. [**17**], участвовали 14 многозначных существительных. Обучающее множество строилось из предложений корпуса новостных текстов San Jose Mercury news (30 млн. слов), содержащих однозначных «родственников» каждого значения по WordNet (приоритет отдавался ближайшим родственникам — синонимам и гиперонимам-словосочетаниям, содержащим целевое слово). Целью эксперимента было сравнить результаты работы классификатора, обученного двумя способами: на корпусе, полученном автоматически с помощью monosemous relatives, и на корпусе, размеченном вручную.

Это оказалось возможно только для 6 существительных (для 8 остальных не оказалось размеченных вручную обучающих данных). В результатах наблюдался заметный разброс: на 4 существительных (*bill, duty, rate, shot*) классификатор, обученный на «родственниках», показал результаты всего лишь на 1–2% ниже, чем обученный на «ручном» корпусе; на двух других заметно уступил (для *line* — 74,7% против 82,6% и для *work* — 65,2% против 75,3%). На оставшихся 8 существительных обученный методом monosemous relatives классификатор достиг наибольшей априорной вероятности или незначительно превзошел ее.

Метод, предложенный Р. Михалца и Д. Молдованом [**18**], основан на использовании не только однозначных синонимов, но и — при их отсутствии — выражений, входящих в толкования значений синсетов (глоссы) в WordNet. Из этих данных формируются поисковые запросы в системе AltaVista (при этом выражения, входящие в толкования, преобразуются в запросы с операторами NEAR, AND, OR; предварительно из них устраняются стоп-слова). Авторы опробовали метод на 20 многозначных словах разных частей речи (в сумме 120 значений). Цель эксперимента состояла в оценке эффективности рассматриваемого метода для построения обучающего множества.

Для ручной проверки качества полученных примеров было отобрано максимум по 10 примеров на каждое значение (всего 1081 пример). В результате проверки 91% примеров были оценены как правильные. Всего же метод позволил получить 80741 пример — для сравнения, в корпусе SemCor [**19**] для тех же 20 слов примеров обнаружилось в тридцать раз меньше — 2581.

В работе Р. Михалца 2002 года [**20**] для порождения большого объема аннотированных текстов использована комбинированная стратегия — однозначные «родственники» плюс глоссы и бутстрэппинг. Начальное множество размеченных примеров (starting seeds) для итеративного алгоритма строится по четырем источникам:

(*1*) Semcor,

(*2*) WordNet,

(*3*) примеры, полученные описанным ранее методом [**18**],

(*4*) аннотированные вручную корпуса (в данном случае использовался корпус SENSEVAL-2).

Используя начальные примеры в качестве запросов для интернет-поиска, алгоритм получает новые размеченные фрагменты, а в них в свою очередь обнаруживает новые примеры для поиска фрагментов. Таким образом был построен корпус GenCor, состоящий из 160 тыс. примеров. Эксперимент по разметке значений нескольких существительных показал, что один и тот же классификатор, будучи обучен на корпусе GenCor и на корпусе, размеченном вручную, получает вполне сопоставимые результаты.

Еще одна альтернатива ручной разметке корпуса — методы, основанные на использовании **псевдослов** (pseudowords) [**21**, **22**]. Изначально псеводслова — это искусственно созданные случайные комбинации двух или нескольких однозначных слов одной части речи. Корпуса с размеченными «значениями» псевдослов полезны для разработки и тестирования методов разрешения омонимии, омографии, распознавания схожих по написанию слов (при OCR), но для общей задачи разрешения полисемии особой ценности не представляют.

В последние годы подход, основанный на псевдословах, стал использоваться для разметки корпусов, предназначенных для разрешения реальной лексической многозначности. Псевдослова становятся семантически мотивированными (semantically aware) [**23**], поскольку моделируют отдельные значения многозначного слова посредством семантически близких им слов.

Возможны разные способы отбора компонентов для псевдослов. Самым естественным вариантом является выбор ближайших однозначных родственников в WordNet — синонимов, гипонимов, гиперонимов, согипонимов. Недостатком этого способа при мелкогранулярной разметке оказывается отсутствие у некоторых значений слова ближайших родственников (в таком понимании), имеющих достаточную для обучения классификатора частоту встречаемости [**24**].

Обойти это препятствие позволяет трактовка родства в WordNet как семантического сходства между синсетами, вычисляемого как структурная значимость вершин графа для заданной вершины: это позволяет расширить область поиска кандидатов; из полученного ранжированного по сходству списка однозначных слов отбирается компонент псевдослова, удовлетворяющий заданному ограничению на минимальную частоту [**23**].

Те же авторы предложили еще один способ нахождения однозначных представителей значений полисемичных слов: в качестве таковых они используют не слова, а автоматически порождаемые тематические сигнатуры (topic signatures [**25**]) — взвешенные тематические вектора, ассоциированные со значениями. Такие вектора могут быть получены на больших корпусах или методом однозначных родственников. Из каждого вектора отбирается однозначное слово с наибольшим весом (т. е. наиболее связанное с заданным значением), имеющее достаточную минимальную частоту. Для 20 существительных был выполнен корелляционный анализ результатов работы классификатора, обученного на размеченном вручную корпусе и на корпусе с автоматически размеченными псевдозначениями. Оценивались три вида псевдослов: построенные методом «семантического сходства», на тематических сигнатурах и случайные псевдослова. Ближе всего к «настоящим» результатам (полученным на размеченных вручную реальных словах) оказались результаты для псевдослов, построенных методом «семантического сходства».

1.3. Гранулярность

Отражаемые в лексических ресурсах лексикографические тонкости оборачиваются чрезмерной дробностью значений слов в инвентаре для WSD, что затрудняет аннотирование текста и служит источником ошибок при WSD. Эту проблему часто называют *WordNet problem*, поскольку в качестве стандарта де-факто для мелкогранулярного (fine-grained) WSD используется WordNet.

Между тем, потребность в таком мелкогранулярном различении значений ограничена узким кругом задач. Один из способов

избежать проблем, вытекающих из излишней детализации значений, состоит в **укрупнении собственных значений слов** в инвентаре. Большинству практических приложений достаточно огрубленного (coarse-grained) распознавания значений, близкого к различению омонимов. Например, 7 значений английского слова *paper* в WordNet можно объединить в три относительно омонимичных значения — (1) материал, (2) статья или письменная работа, (3) газета (предмет, издание, издательство) [**26**].

В иерархически устроенных ресурсах такие укрупненные лексические значения могут быть получены приведением нескольких значений низшего уровня к одному родительскому. В WordNet значения слова не образуют иерархии, поэтому для получения «огрубленной» версии WordNet в задаче coarse-grained WSD (SemEval-7 [**27**]) использовался другой лексикографический ресурс: значения WordNet распределялись по толкованиям верхнего уровня в Oxford Dictionary of English.

Другой вариант крупногранулярной постановки задачи WSD, возникший на стыке собственно WSD и извлечения информации [**28**], можно назвать **семантической категоризацией**. В тезаурусе WordNet каждому синсету приписана довольно общая семантическая категория — *act*, *artifact*, *state*, *substance* и т. п. (всего их 41). Набор этих семантических категорий (*supersenses* [**29**]) и используется в качестве инвентаря значений. Такая постановка задачи — распознавание онтологической категории слова — не заменяет WSD в традиционном понимании, но имеет свою область применения и обладает рядом преимуществ, среди которых относительная простота аннотирования, в том числе несловарной лексики.

2. Специфика задачи различения событийного и несобытийного значений

Задача распознавания событийного и несобытийного употребления неоднозначного существительного представляет собой вариант задачи семантической категоризации, поскольку противопоставление событийного и несобытийного имеет под собой не столько лексическую, сколько онтологическую подоплеку — это противопоставление двух семантических категорий очень высокого уровня (суперкатегорий).

То, что разделение значений многозначного существительного на событийное и несобытийное может не коррелировать ни с какой степенью огрубления его лексических значений, проиллюстрируем на примере значения слова *отделение* в Словаре русского языка [**30**]. Как видно из таблицы 2, несобытийные значения слова не поддаются обобщению до одного лексического значения.

Отметим некоторые особенности задачи WSD в нашей постановке.

Во-первых, событийно-несобытийная многозначность существительного представляет собой пример метонимически мотивированной полисемии (имя ситуации — имена ее участников), в целом относительно далеко отстоящей от омонимии[1] — это хорошо заметно на примерах в таблице 1. Известно, что для автоматического разрешения омонимии успешно применяются тематические признаки. Можно предположить, что, в силу частой тематической общности событийных и несобытийных значений, диагностическая ценность широкого тематического контекста для их различения в общем случае невелика.

Во-вторых, многозначные существительные в событийном и несобытийном значении:

- могут различаться набором активных синтаксических валентностей. Так, у существительных в значении объекта, средства или инструмента отсутствует одноименная валентность, имеющаяся у событийного «омонима»;
- могут различаться синтаксическим способом выражения семантически близких валентностей. Например, одна и та же лексема в творительном падеже может выражать агент при существительном *публикация* в событийном смысле (*публикация автором, издательством*), а в родительном — квазиагент при его несобытийном «омониме» (*публикация автора, издательства*).

Кроме того, событийным «омонимам», в отличие от несобытийных, свойственна сочетаемость с лексико-функциональными глаголами — значениями лексических функций семейства OPER-LABOR-FUNC [31] (*осуществлять управление*, *совершить покупку*, *состоялось отделение*), с различной модальной и околомодальной лексикой (*возможна публикация*, *возможность публикации*, *разрешить публикацию*, *запрет на публикацию*), с темпоральными предлогами (*до*, *после*, *в процессе*).

Все это говорит в пользу значимости для рассматриваемой задачи лексико-синтаксического контекста — как подчиненного, так и подчиняющего.

---

[1]Полисемия (многозначность) и омонимия — относительные понятия, отражающие степень сходства между значениями. Обязательное условие многозначности — наличие «нетривиальной» общей части значений, тогда как крайняя степень омонимии — «чисто внешнее совпадение двух и более слов, в значениях которых нет ничего общего» (Ю. Д. Апресян. Избранные труды. Том 1. Лексическая семантика. Синонимические средства языка. М., 1995).

Таблица 2. Значения слова *отделение* по Словарю русского языка

1. Действие по знач. глаг. *отделить—отделять* и *отделиться—отделяться.*
   Отделение церкви от государства. *Отделение слюны.*

2. Обычно мн. ч. (*отделéния, -ий*). Вещество, выделенное организмом; выделение. *Гнойные отделения.*

3. Отгороженная, обособленная часть помещения.
   Бабушка показала ему переделки в конюшнях, показала и лошадей, и особое отделение для птиц. (И. Гончаров, Обрыв.) *В почтовом поезде, шедшем из Петербурга в Москву, в отделении для курящих, ехал молодой поручик Климов.* (Чехов, Тиф.)
   ‖ Обособленная часть какого-л. вместилища.
   Все отделения пенала Петя наполнил предназначенными для них предметами. (Катаев, Белеет парус одинокий.)
   [*Ящик*] *разгорожен внутри перегородками на отделения — одни широкие, вместительные, другие узкие, глубокие.* (Тендряков, Не ко двору.)

4. Часть учреждения, предприятия и т. п.
   Хирургическое отделение больницы. Литературное отделение факультета.
   □ *Механический цех был разбит на несколько отделений — токарное, револьверное, строгальное, расточное и фрезерное.* (П. Б. Быков, Путь к счастью.)
   ‖ Отдельное учреждение, организация и т. п. в системе чего-л.
   Ленинградское отделение издательства. Отделение милиции. Почтовое отделение связи.

5. Устар. Тематически объединенная часть книги, газеты и т. п.; раздел.
   Журнал сей будет выходить каждый месяц по одной книжке. Каждая книжка будет заключать в себе четыре отделения. (Пушкин, Общество московских литераторов.)

6. Самостоятельная часть концерта, театрального представления и т. п.
   Во втором отделении [*концерта*] *ведущий объявил вокальный номер.* (Первенцев, Честь смолоду.)

7. Низшее воинское подразделение, часть взвода.
   Стрелковое отделение.
   □ *Отделение сержанта Якова Олейника размещалось в большом шалаше.* (Бубеннов, Белая береза.)

Заметим, что в некоторых употреблениях различение событийного и несобытийного значения существительного затруднено или невозможно: например: *пойти за покупками* (‘пойти, чтобы сделать покупки’ или ‘пойти за товарами’), *на стройке стадиона было разрушено более 100 несущих бетонных колонн* (‘в процессе строительства’ или ‘на стройплощадке’).

Встречаются и речевые ошибки, когда смешение событийного и несобытийного значений в одном употреблении некорректно, например:

Публикацию *в «КП» стенограммы разговора под заголовком «Запад вам аплодирует за то, что разваливаете страну» на сайте прочитало несколько сотен тысяч человек.*

В этом примере у существительного *публикация* заполнена валентность на объект (*стенограмма*). Такая валентность есть только у событийного «омонима», однако очевидно, что публикация-событие не может быть объектом прочтения.

Она снимает пальто, одевает свою старенькую бобровую шубку, не раз латанную, велит мне договориться на кассе, чтобы ***покупку*** *отложили на пару часов или до завтра, а сама выходит на улицу.*

Речь в отрывке идет о покупке нового пальто. Отложить покупку (событие) может только сам покупатель. А к кассиру обращаются, чтобы отложить (убрать) на время товар.

3. Метод решения задачи событийно-несобытийной категоризации

В контексте извлечения из текста информации о событиях задача выявления существительных с событийным значением чрезвычайно актуальна, поскольку доля именных упоминаний событий хоть и уступает глагольным, но довольно высока. Так, в обучающем корпусе для ACE 2005 более 35% (1934 из 5349) упоминаний событий приходится на существительные [**32**].

Проблема автоматического различения событийного и несобытийного употребления неоднозначных существительных затрагивается в небольшом числе зарубежных публикаций. Статистические методы, предлагаемые для ее решения, опираются на размеченные корпуса текстов [**33**], [**34**], [**32**].

Для решения поставленной задачи мы также будем использовать методы машинного обучения, зарекомендовавшие себя как наиболее эффективные средства WSD.

В области WSD принято создавать отдельный классификатор для каждой многозначной лексической единицы, так как набор

значений у каждого слова уникален. Несмотря на то, что наша задача требует различения лишь двух классов словоупотреблений (событийного и несобытийного), мы не будем строить единственный обобщенный классификатор, предполагая, что контексты интересующих нас неоднозначных лексических единиц существенно различны и поэтому стратегия «отдельный классификатор для каждой лексической единицы» будет более успешна.

При такой стратегии ключевой проблемой становится подготовка обучающих множеств для каждой многозначной лексической единицы.

Формирование обучающих множеств вручную для столь обширного класса лексических единиц, как «способные к употреблению в событийном и несобытийном значении», невозможно в силу огромной трудоемкости. Даже применение бутстрэппинга, предполагающего небольшой объем ручной разметки для каждой многозначной лексической единицы, представляется затруднительным. Поэтому естественным решением в данных условиях было бы применение подхода, опирающегося на автоматическую разметку обучающего множества. Для автоматического аннотирования обучающего множества мы решили воспользоваться приемом замены неоднозначных существительных однозначными (в смысле событийности-несобытийности), а в качестве основания для подбора замен использовать дистрибутивное сходство.

Рассмотрим подробнее следующие аспекты исследованного нами метода:

- подготовка обучающего множества;
- формирование вектора признаков;
- алгоритм классификации.

3.1. Подготовка обучающего множества

В основе нашего подхода к построению обучающего множества контекстов лежит тот же принцип, на котором построен метод однозначных родственников (monosemous relatives), — а именно принцип замены: каждое из дифференцируемых значений целевого многозначного слова заменяется набором специально подобранных однозначных слов.

Однозначность в нашем случае понимается в соответствии с решаемой задачей — как однозначная принадлежность слова-замены (во всех его возможных употреблениях) к категории либо события, либо несобытия.

Слова-замены отбираются из существительных, дистрибутивно близких целевому. В нашем эксперименте используется модель дистрибутивной семантики, основанная на синтаксических отношениях [35].

Для опробования метода были выбраны шесть контрольных существительных, неоднозначных с точки зрения событийности/несобытийности: *организация, отделение, отопление, публикация, стройка, управление.*

Для каждого контрольного слова (за исключением слова *отделение*) из топ-100 близких ему слов в дистрибутивной модели было отобрано до 10 слов-замен, имеющих только событийные значения, и до 10 слов, не имеющих событийных значений. Отбор выполнялся вручную. При этом предпочтение отдавалось словам, концептуально (онтологически) наиболее схожим с целевым и более частотным. Для слова *отделение* в верхней сотне и даже в топ-500 дистрибутивно близких слов не нашлось подходящих (хоть сколько-нибудь концептуально близких) существительных с событийным значением. В связи с этим из тезауруса `RuWordNet`[2] были взяты два синонима слова *отделение* — *отсоединение* и *разъединение*, а затем из списка дистрибутивных соседей этих слов были выбраны слова-замены с событийным значением.

Списки замен для контрольных слов приведены в таблице 3.

Процедура получения аннотированного корпуса для каждого целевого неоднозначного существительного W состояла в следующем. В массиве новостных текстов отыскивались предложения, содержащие событийные или несобытийные замены для существительного W. Эти предложения копировались в обучающее множество, а каждое употребление слова-замены аннотировалось соответствующим ему категориальным значением существительного W: W-событийное или W-несобытийное.

3.2. Подходы к формированию вектора признаков

В задаче классификации событийных и несобытийных употреблений многозначного существительного большей дифференцирующей силой должны обладать локальные признаки, нежели тематические. Зачастую широкий тематический контекст у обоих вариантов употребления многозначного слова оказывается общим (например, у слов *покупка, стройка, публикация* и многих других). Поэтому в своих экспериментах мы ограничились рассмотрением только локальных признаков. Было исследовано два подхода к формированию вектора признаков.

[2]http://ruwordnet.ru/ru/

Таблица 3. Слова-замены для контрольных неоднозначных существительных

Контрольное существительное	Событийные замены	Несобытийные замены
организация	*деятельность, агитация, демонстрация, митинг, заговор, манифестация*	*партия, структура, предприятие, компания, фирма, корпорация, профсоюз, банк, комитет, правительство*
отделение	*выбрасывание, приостановление, выталкивание, отключение, блокирование, выключение, роспуск, приостановка, сворачивание, отслоение*	*филиал, стационар, отдел, факультет, округ, поликлиника, кафедра, мэрия, администрация, комитет*
отопление	*теплоснабжение, электроснабжение, теплопередача, энергообеспечение, газоснабжение, подогрев, энергоснабжение, нагрев, кондиционирование, водоотведение*	*обогреватель, конвектор, котельная, термостат, кондиционер, нагреватель, электронагреватель, калорифер, холодильник, камин*
публикация	*перепубликация, обнародование, ратификация, выдвижение, подписание, предоставление, проведение, оглашение, цитирование, опубликование*	*стенограмма, монография, мемуары, сборник, заметка, брошюра, фельетон, эссе, статейка, очерк*
стройка	*возведение, строительство, реформирование, реорганизация, реставрация, снос, модернизация, реструктуризация*	*стройплощадка, недострой, новостройка, высотка, пустырь, многоэтажка*
управление	*планирование, регулирование, обеспечение, бюджетирование, администрирование, подчинение, принуждение, снабжение, подключение, сопровождение, пилотирование*	*отдел, ведомство, департамент, бюро, комитет, училище, филиал, предприятие, дирекция, министерство*

Первый вариант в качестве признаков экземпляра (употребления многозначного слова) использовал нормальные формы слов, находящихся в его линейной окрестности. Размер окна составлял ±3 слова. Позиция слов контекста относительно многозначного слова не учитывалась. Исключение составляли лишь предлоги — для них в признак включалась информация о том, слева или справа от многозначного слова данный предлог был употреблен. В дальнейшем данный метод извлечения признаков будем называть **окно-3**.

Второй вариант опирался на «синтаксическое окно» (древесная окрестность). В него включались слова, находящиеся в каком-либо синтаксическом отношении с многозначным словом. Признак представлял собой тройку:

- направление синтаксической связи;
- тип синтаксической связи;
- нормальная форма связанного слова.

В случае если синтаксическая связь вела к предлогу, признак дополнялся нормальной формой слова по ту сторону предлога и (с учетом типа синтаксической связи) становился пятикомпонентным. Кроме того, для таких пятикомпонентных признаков формировался признак-дублер, в котором предлог абстрагировался[3]. Данный метод в дальнейшем будет упоминаться как **синтаксис**.

Рассмотрим на примере, какие наборы признаков формируются каждым из методов. Пусть дана следующая фраза, содержащая многозначное слово *отделение*.

Сержант был госпитализирован в реанимационное отделение местной больницы.

Для нее вышеупомянутые методы извлечения признаков дадут следующие результаты (табл. 4).

Поскольку в основе системы признаков лежат лексические единицы, совокупный вектор признаков очень велик. Для ограничения размерности вектора применялся отбор признаков по критерию χ^2. Известно, что данный критерий является одним из наиболее эффективных при отборе признаков в лексическом признаковом пространстве [**36**].

Напомним, что рассматривается двуклассовая задача классификации (событийное/несобытийное употребление). Для каждого признака F и каждого класса C рассмотрим бинарные случайные величины на множестве обучающих примеров S, которые мы будем обозначать

[3]Прием с абстрагированием предлогов нацелен на решение проблемы с различием моделей управления многозначного слова и его замен.

Таблица 4. Примеры признаков словоупотребления

окно-3	**синтаксис**
госпитализировать в (слева) реанимационный местный больница	‹вниз, определительная, реанимационный› ‹вниз, квазиагентивная, больница› ‹вверх, предложная, в, 2-комплетивная, госпитализировать› ‹вверх, предложная, $*$, 2-комплетивная, госпитализировать›

теми же буквами F и C. Случайная величина F принимает значение 1 на обучающем примере s, если соответствующий признак присутствует в s, и принимает значение 0 в противном случае. Случайная величина C принимает на обучающем примере s значение 1, если s принадлежит классу C, и принимает значение 0 в противном случае.

Выдвинем гипотезу о независимости случайных величин F и C. Введем следующее обозначение для их теоретического совместного распределения:

$$p(f,c) = P(F = f, C = c) = P(F = f)P(C = c), \quad \text{где} \quad f, c \in \{0,1\}.$$

В то же время эмпирическая функция совместного распределения имеет вид:

$$\hat{p}(f,c) = \frac{|\{s \in S \mid F(s) = f,\ C(s) = c\}|}{|S|}$$

Рассмотрим статистику

$$\chi^2(F,C) = |S| \sum_{f \in \{0,1\}} \sum_{c \in \{0,1\}} \frac{(\hat{p}(f,c) - p(f,c))^2}{p(f,c)},$$

которая при $|S| \to \infty$ распределена по закону χ^2 с тремя степенями свободы, при условии, что гипотеза о независимости случайных величин F и C верна.

Согласно критерию Пирсона, чтобы оценить вероятность справедливости выдвинутой гипотезы, можно воспользоваться квантилями распределения χ^2 с тремя степенями свободы. Например, при $\chi^2(F,C) > 11.345$ вероятность справедливости гипотезы меньше 0.01, следовательно, вероятность зависимости случайных величин F и C больше 0.99. Таким образом, при больших значениях рассматриваемой статистики выдвинутая гипотеза должна быть отклонена. Это означает зависимость признака F и класса C, а следовательно F обладает классифицирующей силой.

Для исключения неинформативных признаков квантили распределения χ^2 не использовались, вычислялась лишь величина $\chi^2(F, C)$.

На практике использовался следующий способ ее вычисления:

$$\chi^2(F, C) = \frac{(N_{0,0} + N_{0,1} + N_{1,0} + N_{1,1})(N_{0,0}N_{1,1} - N_{0,1}N_{1,0})}{(N_{0,0} + N_{0,1})(N_{1,0} + N_{1,1})(N_{0,0} + N_{1,0})(N_{0,1} + N_{1,1})},$$

где $N_{f,c} = |S|\hat{p}(f, c)$ обозначает количество обучающих примеров s, для которых $F(s) = f$ и $C(s) = c$.

За итоговую значимость признака F был принят максимум величины $\chi^2(F, C)$ по всем классам. Все признаки ранжировались по степени значимости (чем больше $\max\limits_{C} \chi^2(F, C)$, тем лучше признак F) и отбиралось фиксированное количество наилучших признаков (даже если не все они достаточно «надежны»). В дальнейшем такой отбор фиксированного количества признаков будет именоваться «подрезанием по порогу».

Поскольку примеры в обучающее множество отбирались по принципу замены, то среди слов, составляющих контекст аннотированного слова, оказывались слова, которые не встречаются вблизи исходного многозначного слова (для которого строится классификатор). Поэтому для очистки вектора признаков использовался следующий прием. Из представительного корпуса текстов извлекались примеры употребления исходного многозначного слова. Для каждого примера извлекались признаки. Их совокупность формировала вектор таких признаков, которые фактически встречаются в окрестности многозначного слова. За результирующий вектор признаков бралось пересечение этого вектора с вектором признаков, полученным по обучающему множеству (с аннотированными словами-заменами).

3.3. Алгоритм классификации

В качестве алгоритма классификации мы использовали наивный байесовский классификатор (НБК). Так как вероятность класса[4] нам не известна, мы исключили из формулы соответствующий множитель, сделав допущение о равноверятности классов.

В расчётах использовался полный вектор признаков. Если у классифицируемого примера признак отсутствовал, то множителем считалась вероятность $p(\rceil F_i|C)$.

[4]То есть информация о том, как часто многозначное слово употребляется в событийном и, наоборот, в несобытийном значении.

Таблица 5. Размеры обучающих множеств в экспериментах с методом **окно-3 + НБК**

Многозначное слово	Примеров несобытийного значения	Примеров событийного значения
организация	2 906 363	377 780
отделение	855 105	26 832
отопление	16 351	23 550
публикация	15 930	421 821
стройка	13 483	461 605
управление	1 435 964	103 546

Технически мы применяли вариант классификатора, выраженного через сумму логарифмов. Результирующая формула имеет вид:

$$class = \arg\max_{c \in C} \sum_{i=1}^{n} \log p(F_i|c).$$

Если $p(F_i|c)$ равнялся нулю, то в качестве значения использовалось 10^{-16}.

4. Эксперименты

4.1. Описание эксперимента

Оценка эффективности исследованных нами методов выполнялась на 6 контрольных словах.

Примеры употреблений слов-замен в контексте извлекались из корпуса БДТФ-Н[5]. Корпус содержит более 34 млн. уникальных записей (примерно соответствующих предложениям), полученных из 3,7 млн. новостных сообщений. Предварительно для корпуса был выполнен морфологический и синтаксический анализ посредством инструментария, описанного в [**37**]; поиск слов-замен выполнялся по нормальной форме. В результате для каждого контрольного многозначного слова было сформировано обучающее множество, состоящее из предложений, разделенных на два класса (содержащих событийные и несобытийные замены). В таблицах 5 и 6 указаны размеры обучающих множеств и соотношение примеров событийного и несобытийного класса.

[5]http://ai-center.botik.ru/Airec/index.php/ru/databases/39-dbtf-n

Таблица 6. Размеры обучающих множеств в экспериментах с методом **синтаксис + НБК**

Многозначное слово	Примеров несобытийного значения	Примеров событийного значения
организация	2 786 504	364 545
отделение	832 648	24 867
отопление	11 783	20 443
публикация	14 428	403 631
стройка	11 214	425 084
управление	1 388 367	127 046

Таблица 7. Размеры вектора признаков

Многозначное слово	Количество признаков	
	окно-3 + НБК	**синтаксис + НБК**
организация	27 769	31 191 (30 000)
отделение	12 805	10 159
отопление	2 476	768
публикация	7 814	6 095
стройка	5 064	2 993
управление	19 965	19 477

Тестовые множества[6] (примеры употреблений контрольных существительных в событийном и несобытийном значении) подготавливались вручную. В качестве источника послужили новостные сюжеты из интернета. Размеры тестовых множеств для всех контрольных слов одинаковы — по 100 примеров каждого класса.

Размер вектора признаков в экспериментах подрезался по порогу 30000. Фактические размеры векторов признаков приведены в таблице 7 (в скобках указан размер подрезанного вектора).

4.2. Результаты

Результаты разрешения событийно-несобытийной неоднозначности на контрольном множестве приведены в таблице 8. Основным показателем эффективности WSD-метода выступает точность (accuracy) классификации значений. Дополнительно в скобках указано покрытие

[6] http://doi.org/10.5281/zenodo.1308734

Таблица 8. Результаты разрешения неоднозначности в формате «accuracy (coverage)», %

Многозначное слово	**синтаксис + НБК**	**окно-3 + НБК**
организация	67.70 (96)	60.0 (100)
отделение	80.33 (89)	72.0 (100)
отопление	68.14 (67.5)	66.33 (99.5)
публикация	79.57 (93.5)	61.8 (99.5)
стройка	77.19 (85.5)	75.37 (99.5)
управление	91.37 (98.5)	79.5 (100)

(coverage) решения (процент классифицированных[7] примеров тестового множества).

Эксперименты на небольшой выборке в целом подтвердили работоспособность метода, основанного на дистрибутивных заменах, для решения событийно-несобытийной неоднозначности существительных. Результаты согласуются с выдвинутым предположением о значимости именно синтаксического контекста для решения поставленной задачи. На всех контрольных многозначных словах синтаксический подход превзошел линейный.

Подводя итоги, обозначим направления исследований, которые, на наш взгляд, выглядят перспективными для дальнейшего развития описанного метода разрешения событийно-несобытийной неоднозначности.

(*1*) Наиболее слабым компонентом нашего метода оказалась модель дистрибутивной семантики, использованная для подбора слов-замен. Эта модель, хоть и построена на основе сходства лексико-синтаксического контекста (что само по себе хорошо для задачи разрешения категориальной неоднозначности), но учитывает всего три типа синтаксических отношений (тогда как алгоритм разрешения неоднозначности опирается на все множество типов синтаксических связей). Естественно было бы ожидать, что чем больше типов синтаксических отношений будет положено в основу дистрибутивного сходства, тем более подходящими для конечной задачи окажутся замены целевого слова на дистрибутивно близкие.

(*2*) При отборе слов-замен процедура их оценки с точки зрения событийно-несобытийной неоднозначности выполнялась человеком.

[7]Тестовый пример не подвергается классификации, если у него не обнаружилось ни одного признака, входящего в вектор отобранных признаков.

Автоматизация этой процедуры безусловно повысила бы практическую ценность метода. Сделать это можно следующим образом. Предварительно вручную сформировать два эталонных множества частотных существительных — таких, которые имеют только событийные значения, и таких, которые не имеют событийных значений. Для каждого кандидата на роль слова-замены оценить степень близости к эталонным множествам и отбирать только те, которые явно тяготеют к одному из них.

Заключение

Изложенный в статье подход к разрешению событийно-несобытийной неоднозначности существительных разрабатывался как составная часть метода извлечения событийной информации из текстов. Одним из ключевых требований было минимизировать трудозатраты на построение классификаторов из-за необходимости работать с обширным пластом многозначной лексики. Для автоматического аннотирования обучающего корпуса мы использовали оригинальный метод однозначных замен по принципу дистрибутивного сходства. Результаты, полученные на небольшой выборке контрольных слов, в целом можно считать обнадеживающими (средняя точность по выборке составляет 77,38% при классификации на основе синтаксического контекста, хотя разброс значений для разных слов составляет более 20%).

Результаты показали значимость синтаксических признаков контекста при разрешении событийно-несобытийной неоднозначности. На всех многозначных существительных, отобранных для экспериментального исследования, использование синтаксического контекста дало лучшие результаты по сравнению с линейным.

Дальнейшее развитие метода мы видим, во-первых, в построении и использовании более совершенной модели дистрибутивной семантики, а во-вторых — в полной автоматизации процедуры подбора однозначных слов-замен.

Список литературы

[1] Р. И. Гусейнов. «О взаимодействии ономасиологии и грамматики», *Вопросы языкознания*, **6** (1987), с. 65–71. URL ↑4

[2] А. Г. Пазельская. «Аспектуальность и русские предикатные имена», *Вопросы языкознания*, **4** (2003), с. 72–90. URL ↑4

[3] Е. В. Падучева. «Предикатные имена в лексикографическом аспекте», *НТИ. Сер. 2*, 1991, №5, 18 с. URL ↑4

[4] Е. А. Лютикова. «Стратегии оформления аргументов в русской событийной номинализации и теория падежа», *Труды Института русского языка им. В. В. Виноградова*, **10** (2016), с. 201–220. ↑4

[5] Ю. Д. Апресян, *Избранные труды.* Т. 1: *Лексическая семантика. Синонимические средства языка*, Язык. Семиотика. Культура, 2-е изд., перераб. и доп., Языки славянской культуры, М., 1995, 481 с. ↑4,5

[6] W. Weaver. “Translation”, *Machine translation of languages: Fourteen essays*, eds. W. N. Locke, A. D. Booth, Technology Press of MIT, Cambridge, MA; John Wiley & Sons, New York, NY, 1955, pp. 15–23. ↑6

[7] J. Hutchins. “First steps in mechanical translation”, Proceedings of MT Summit VI: Past, Present, Future (San Diego, CA, 1997). URL ↑6

[8] R. Navigli. “Word sense disambiguation: a survey”, *ACM Computing Surveys*, **41**:2 (2009), 10, 69 pp. URL doi ↑7

[9] R. Navigli. “A quick tour of word sense disambiguation, induction and related approaches”, SOFSEM 2012: Theory and Practice of Computer Science, Lecture Notes in Computer Science, vol. **7147**, eds. M. Bielikova, G. Friedrich, G. Gottlob, S. Katzenbeisser, G. Turan, Springer, Berlin–Heidelberg, 2012, pp. 115–129. URL doi ↑7

[10] D. Yarowsky. “Unsupervised word sense disambiguation rivaling supervised methods”, *ACL ’95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1995, pp. 189–196. URL doi ↑8

[11] R. Mihalcea. “Co-training and self-training for word sense disambiguation”, *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL*, 2004. URL ↑8

[12] A. Fujii, K. Inui, T. Tokunaga, H. Tanaka. “Selective sampling for example-based word sense disambiguation”, *Computational Linguistics*, **24**:4 (1998), pp. 573–598. URL ↑8

[13] R. Mihalcea, T. Chklovski. “Open mind word expert: creating large annotated data collections with web users’ help”, *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora*, LINC 2003 (Budapest, Hungary, 2003), pp. 53–60. URL ↑8

[14] W. Gale, K. W. Church, D. Yarowsky. “A method for disambiguating word senses in a large corpus”, *Computers and the Humanities*, **26**:5/6 (1992), pp. 415–439. doi ↑8

[15] D. Yarowsky. “Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora”, *COLING ’92 Proceedings of the 14th conference on Computational linguistics.* V. 2, 1992, pp. 454–460. URL doi ↑8

[16] C. Leacock, M. Chodorow. “Combining local context with WordNet similarity for word sense identification”, *WordNet: A Lexical Reference System and its Application* (Cambridge, MA), ed. Ch. Fellbaum, MIT Press, 1998, pp. 265–283. URL ↑9

[17] C. Leacock, M. Chodorow, G. Miller. “Using corpus statistics and WordNet

relations for sense identification", *Computat. Ling.*, **24**:1 (1998), pp. 147–165. URL ↑9

[18] R. Mihalcea, D. Moldovan. "An automatic method for generating sense tagged corpora", Proc. of AAAI 99 (Orlando, FL, 1999). URL ↑9,10

[19] G. A. Miller, M. Chodorow, S. Landes, C. Leacock, R. G. Thomas. "Using a semantic concordance for sense identification", *HLT '94 Proceedings of the workshop on Human Language Technology*, ARPA, 1994, pp. 240–243. URL doi ↑10

[20] R. Mihalcea. "Bootstrapping large sense tagged corpora", *Proceedings of the Third International Conference on Language Resources and Evaluation*, LREC-2002 (Las Palmas, Canary Islands, Spain, 2002), ELRA. URL ↑10

[21] W. Gale, K. Church, D. Yarowsky. "On evaluation of word-sense disambiguation systems", 30th Annual Meeting of the Association for Computational Linguistics (Columbus, OH, 1992), pp. 249–256. ↑10

[22] D. Yarowsky. "One sense per collocation", *HLT'93: Proceedings of the workshop on Human Language Technology* (Morristown, NJ, USA, 1993), Association for Computational Linguistics, pp. 266–271. URL doi ↑10

[23] M. T. Pilehvar, R. Navigli. "A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation", *Computational Linguistics*, **40**:4 (2014), pp. 837–881, J14-4005. URL doi ↑10,11

[24] L. Otrusina, P. Smrz. "A new approach to pseudoword generation", *Proceedings of the Seventh conference on International Language Resources and Evaluation*, LREC'10 (Valletta, Malta, 2010), pp. 1195–1199. URL ↑11

[25] Ch.-Y. Lin, E. Hovy. "The automated acquisition of topic signatures for text summarization", *COLING '00 Proceedings of the 18th conference on Computational linguistics*. V. 1 (Saarbrucken, 2000), pp. 495–501. doi ↑11

[26] N. Ide, Y. Wilks. "Making sense about sense", Word Sense Disambiguation: Algorithms and Applications (Dordrecht, The Netherlands), 2006, pp. 47–74. ↑12

[27] R. Navigli, K. C. Litkowski, O. Hargraves. "Semeval-2007 task 07: Coarse-grained english all-words task", *SemEval '07 Proceedings of the 4th International Workshop on Semantic Evaluations* (Prague, 2007), pp. 30–35. ↑12

[28] N. Schneider, D. Hovy, A. Johannsen, M. Carpuat. "SemEval-2016 task 10: detecting minimal semantic units and their meanings (DiMSUM)", Proceedings of SemEval-2016 (San Diego, California, 2016), pp. 546–559. URL ↑12

[29] M. Ciaramita, Y. Altun. "Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger", *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, 2006), pp. 594–602. URL ↑12

[30] А. П. Евгеньева (ред.). *Словарь русского языка*, В 4-х т., РАН, Ин-т лингвистич. исследований 4-е изд., стер., Рус. яз., М.; Полиграфресурсы, 1999. ↑12

[31] Ю. Д. Апресян. «О семантической непустоте и мотивированности глагольных лексических функций», *Вопросы языкознания*, **4** (2004), с. 3–18. URL ↑13

[32] C. Creswell, M. J. Beal, J. Chen, Th. L. Cornell, L. Nilsson, R. K. Srihari. “Automatically Extracting Nominal Mentions of Events with a Bootstrapped Probabilistic Classifier”, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (Sydney, Australia), 2006, pp. 168–175 (English). URL ↑15

[33] R. Saurí, R. Knippen, M. Verhagen, J. Pustejovsky. “Evita: A Robust Event Recognizer For QA Systems”, HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (Vancouver, British Columbia, Canada), 2005, pp. 700–707 (English). URL ↑15

[34] I. Russo, T. Caselli, F. Rubino. “Recognizing deverbal events in context”, *International Journal of Computational Linguistics and Applications*, **2**:1-2 (2011), pp. 91 (English). ↑15

[35] I. V. Trofimov, E. A. Suleymanova. “A syntax-based distributional model for discriminating between semantic similarity and association”, Proceedings of the International Conference “Dialogue 2017”, *Computational Linguistics and Intellectual Technologies*, **1**:16 (2017), pp. 349–359. URL ↑17

[36] Y. Yang, J. O. Pedersen. “A Comparative Study on Feature Selection in Text Categorization”, Proceedings of the Fourteenth International Conference on Machine Learning, ICML ’97 (San Francisco, CA, USA), 1997, pp. 412–420 (English). URL ↑19

[37] S. Sharoff, J. Nivre. “The proper place of men and machines in language technology: processing Russian without any linguistic knowledge”, Papers from the Annual International Conference “Dialogue”, *Computational Linguistics and Intellectual Technologies*, 2011, pp. 591–604. URL ↑22

Поступила в редакцию 27.07.2018
Переработана 19.10.2018
Опубликована 01.11.2018

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Пример ссылки на эту публикацию:

И. В. Трофимов, Е. А. Сулейманова, Н. А. Власова, А. В. Подобряев. «Разрешение событийно-несобытийной неоднозначности существительных». *Программные системы: теория и приложения*, 2018, **9**:4(39), с. 3–33.

doi 10.25209/2079-3316-2018-9-4-3-33

URL http://psta.psiras.ru/read/psta2018_4_3-33.pdf

Об авторах:

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования

0000-0002-6903-4730
e-mail: itrofimov@gmail.com

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, одна из разработчиков технологии построения систем извлечения информации

0000-0002-0792-9651
e-mail: yes2helen@gmail.com

Наталья Александровна Власова

Младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации

0000-0002-7843-6870
e-mail: nathalie.vlassova@gmail.com

Алексей Владимирович Подобряев

Младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, один из разработчиков технологии построения систем извлечения информации

0000-0002-4493-998X
e-mail: alex@alex.botik.ru

CSCSTI 16.31.21, 20.19.27
UDC 004.89:004.912

Igor Trofimov, Elena Suleymanova, Natalia Vlasova, Alexey Podobryaev. *Disambiguation between eventive and non-eventive meaning of nouns.*

Abstract. Event extraction is an advanced form of text mining having numerous applications. One of the challenges faced by event extraction systems is the problem of automatic distinguishing between eventive and non-eventive use of ambiguous event nominals. The proposed disambiguation method relies on an automatically generated training set. In order to learn the difference between eventive and non-eventive reading of a target ambiguous nominal, the classifier is trained on two sets of automatically labelled examples featuring unambiguous distributionally similar lexical substitutes for either reading. The method was evaluated on a small sample of 6 ambiguous event-denoting nouns and performed fairly well (77,38% average accuracy, although with more than 20% variation for individual nouns). Suggestions for future work include development of a more advanced distributional model and research towards automated selection of unambiguous substitutes.

Key words and phrases: word sense disambiguation, automatic training set generation, distributional semantic model, event, event nominal, event-related information extraction.

2010 *Mathematics Subject Classification*: 68T50; 91F20

© I. V. Trofimov, E. A. Suleymanova, N. A. Vlasova, A. V. Podobryaev, 2018
© Ailamazyan Program Systems Institute of RAS, 2018
© Program Systems: Theory and Applications (design), 2018
10.25209/2079-3316-2018-9-4-3-33

References

[1] R. I. Guseynov. “On interaction between onomasiology and grammar”, *Voprosy yazykoznaniya*, **6** (1987), pp. 65–71. URL ↑4

[2] A. G. Pazel’skaya. “Aspectuality and Russian predicate names”, *Voprosy yazykoznaniya*, **4** (2003), pp. 72–90. URL ↑4

[3] Ye. V. Paducheva. “Predicate names in the lexicographic aspect”, *NTI. Ser. 2*, 1991, no.5, 18 pp. URL ↑4

[4] Ye. A. Lyutikova. “Argument-marking strategies in Russian event nominalization and the case theory”, *Trudy Instituta russkogo yazyka im. V. V. Vinogradova*, **11** (2016), pp. 201–220. % ↑4

[5] Yu. D. Apresyan, *Selected works*. V. 1: *Lexical semantics. Synonymous means of language*, Yazyk. Semiotika. Kul’tura, 2-ye izd., pererab. i dop., Yazyki slavyanskoy kul’tury, M., 1995, 481 pp. ↑4,5

[6] W. Weaver. “Translation”, *Machine translation of languages: Fourteen essays*, eds. W. N. Locke, A. D. Booth, Technology Press of MIT, Cambridge, MA; John Wiley & Sons, New York, NY, 1955, pp. 15–23. ↑6

[7] J. Hutchins. “First steps in mechanical translation”, Proceedings of MT Summit VI: Past, Present, Future (San Diego, CA, 1997). URL ↑6

[8] R. Navigli. “Word sense disambiguation: a survey”, *ACM Computing Surveys*, **41**:2 (2009), 10, 69 pp. URL doi ↑7

[9] R. Navigli. “A quick tour of word sense disambiguation, induction and related approaches”, SOFSEM 2012: Theory and Practice of Computer Science, Lecture Notes in Computer Science, vol. **7147**, eds. M. Bielikova, G. Friedrich, G. Gottlob, S. Katzenbeisser, G. Turan, Springer, Berlin–Heidelberg, 2012, pp. 115–129. URL doi ↑7

[10] D. Yarowsky. “Unsupervised word sense disambiguation rivaling supervised methods”, *ACL ’95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1995, pp. 189–196. URL doi ↑8

[11] R. Mihalcea. “Co-training and self-training for word sense disambiguation”, *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL*, 2004. URL ↑8

[12] A. Fujii, K. Inui, T. Tokunaga, H. Tanaka. “Selective sampling for example-based word sense disambiguation”, *Computational Linguistics*, **24**:4 (1998), pp. 573–598. URL ↑8

[13] R. Mihalcea, T. Chklovski. “Open mind word expert: creating large annotated data collections with web users’ help”, *Proceedings of the EACL 2003 Workshop on Linguistically Annotated Corpora*, LINC 2003 (Budapest, Hungary, 2003), pp. 53–60. URL ↑8

[14] W. Gale, K. W. Church, D. Yarowsky. “A method for disambiguating word senses in a large corpus”, *Computers and the Humanities*, **26**:5/6 (1992), pp. 415–439. doi ↑8

[15] D. Yarowsky. “Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora”, *COLING ’92 Proceedings of the 14th conference on Computational linguistics*. V. 2, 1992, pp. 454–460. URL doi ↑8

[16] C. Leacock, M. Chodorow. “Combining local context with WordNet similarity for word sense identification”, *WordNet: A Lexical Reference System and its*

Application (Cambridge, MA), ed. Ch. Fellbaum, MIT Press, 1998, pp. 265–283. ↑9

[17] C. Leacock, M. Chodorow, G. Miller. "Using corpus statistics and WordNet relations for sense identification", *Computat. Ling.*, **24**:1 (1998), pp. 147–165. ↑9

[18] R. Mihalcea, D. Moldovan. "An automatic method for generating sense tagged corpora", Proc. of AAAI 99 (Orlando, FL, 1999). ↑9,10

[19] G. A. Miller, M. Chodorow, S. Landes, C. Leacock, R. G. Thomas. "Using a semantic concordance for sense identification", *HLT '94 Proceedings of the workshop on Human Language Technology*, ARPA, 1994, pp. 240–243. ↑10

[20] R. Mihalcea. "Bootstrapping large sense tagged corpora", *Proceedings of the Third International Conference on Language Resources and Evaluation*, LREC-2002 (Las Palmas, Canary Islands, Spain, 2002), ELRA. ↑10

[21] W. Gale, K. Church, D. Yarowsky. "On evaluation of word-sense disambiguation systems", 30th Annual Meeting of the Association for Computational Linguistics (Columbus, OH, 1992), pp. 249–256. ↑10

[22] D. Yarowsky. "One sense per collocation", *HLT'93: Proceedings of the workshop on Human Language Technology* (Morristown, NJ, USA, 1993), Association for Computational Linguistics, pp. 266–271. ↑10

[23] M. T. Pilehvar, R. Navigli. "A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation", *Computational Linguistics*, **40**:4 (2014), pp. 837–881, J14-4005. ↑10,11

[24] L. Otrusina, P. Smrz. "A new approach to pseudoword generation", *Proceedings of the Seventh conference on International Language Resources and Evaluation*, LREC'10 (Valletta, Malta, 2010), pp. 1195–1199. ↑11

[25] Ch.-Y. Lin, E. Hovy. "The automated acquisition of topic signatures for text summarization", *COLING '00 Proceedings of the 18th conference on Computational linguistics*. V. 1 (Saarbrucken, 2000), pp. 495–501. ↑11

[26] N. Ide, Y. Wilks. "Making sense about sense", Word Sense Disambiguation: Algorithms and Applications (Dordrecht, The Netherlands), 2006, pp. 47–74. ↑12

[27] R. Navigli, K. C. Litkowski, O. Hargraves. "Semeval-2007 task 07: Coarse-grained english all-words task", *SemEval '07 Proceedings of the 4th International Workshop on Semantic Evaluations* (Prague, 2007), pp. 30–35. ↑12

[28] N. Schneider, D. Hovy, A. Johannsen, M. Carpuat. "SemEval-2016 task 10: detecting minimal semantic units and their meanings (DiMSUM)", Proceedings of SemEval-2016 (San Diego, California, 2016), pp. 546–559. ↑12

[29] M. Ciaramita, Y. Altun. "Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger", *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, 2006), pp. 594–602. ↑12

[30] A. P. Eugen'yeva (ed.). *Dictionary of the Russian language in 4 volumes*, in 4 volumes., RAN, In-t lingvistich. issledovaniy 4-ye izd., ster., Rus. yaz., M.; Poligrafresursy, 1999. ↑12

[31] Yu. D. Apresyan. "On semantic non-emptiness and motivatedness of verbal lexical functions", *Voprosy yazykoznaniya*, **4** (2004), pp. 3–18. ↑13

[32] C. Creswell, M. J. Beal, J. Chen, Th. L. Cornell, L. Nilsson, R. K. Srihari. "Automatically Extracting Nominal Mentions of Events with a Bootstrapped

Probabilistic Classifier", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (Sydney, Australia), 2006, pp. 168–175 (English). URL ↑15

[33] R. Saurí, R. Knippen, M. Verhagen, J. Pustejovsky. "Evita: A Robust Event Recognizer For QA Systems", HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (Vancouver, British Columbia, Canada), 2005, pp. 700–707 (English). URL ↑15

[34] I. Russo, T. Caselli, F. Rubino. "Recognizing deverbal events in context", *International Journal of Computational Linguistics and Applications*, **2**:1-2 (2011), pp. 91 (English). ↑15

[35] I. V. Trofimov, E. A. Suleymanova. "A syntax-based distributional model for discriminating between semantic similarity and association", Proceedings of the International Conference "Dialogue 2017", *Computational Linguistics and Intellectual Technologies*, **1**:16 (2017), pp. 349–359. URL ↑17

[36] Y. Yang, J. O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97 (San Francisco, CA, USA), 1997, pp. 412–420 (English). URL ↑19

[37] S. Sharoff, J. Nivre. "The proper place of men and machines in language technology: processing Russian without any linguistic knowledge", Papers from the Annual International Conference "Dialogue", *Computational Linguistics and Intellectual Technologies*, 2011, pp. 591–604. URL ↑22

Sample citation of this publication:

Igor Trofimov, Elena Suleymanova, Natalia Vlasova, Alexey Podobryaev. "Disambiguation between eventive and non-eventive meaning of nouns". *Program Systems: Theory and Applications*, 2018, **9**:4(39), pp. 3–33. (*In Russian*).

DOI 10.25209/2079-3316-2018-9-4-3-33

URL `http://psta.psiras.ru/read/psta2018_4_3-33.pdf`