О.А. Митрофанова, О.Н. Ляшевская, П.В. Паничева

ЭКСПЕРИМЕНТЫ ПО СТАТИСТИЧЕСКОМУ РАЗРЕШЕНИЮ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ РУССКИХ ИМЁН СУЩЕСТВИТЕЛЬНЫХ В КОРПУСЕ[1]

1. Введение

В данной статье рассматриваются новые результаты, полученные в ходе исследований по автоматизации процесса разрешения лексико-семантической неоднозначности текстов[2]. Ранее были проведены серии экспериментов по автоматическому разрешению неоднозначности контекстов употребления предметных имён существительных с различной семантической структурой. Были получены обширные экспериментальные данные на русскоязычном материале и определены оптимальные условия, обеспечивающие достаточно высокое качество разрешения семантической неоднозначности слов в контекстах (от 85% и выше). Оптимальными признаны следующие условия разрешения лексико-семантической неоднозначности в контекстах: а) высокий объём экспериментальной выборки; б) наличие в выборке не менее 100 контекстов употребления слова в исследуемом значении; в) объём эталонного класса около 500 контекстов; оценка бли-

[1] Работа выполнена при частичной финансовой поддержке РФФИ (проект 06-06-80251).

[2] *Митрофанова О.А., Паничева П.В., Ляшевская О.Н.* Статистическое разрешение лексико-семантической неоднозначности в контекстах для предметных имён существительных // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». М., 2008. С. 368–375; *Mitrofanova O., Panicheva P., Lashevskaya O.* Statistical Word Sense Disambiguation in Contexts for Russian Nouns Denoting Physical Objects // Text, Speech and Dialogue. Proceedings of the 11th International Conference TSD 2008, Brno, Czech Republic, September 8–12, 2008. Springer-Verlag, 2008 P. 153–159.

зости контекстов к эталонному классу с использованием значения косинуса угла между контекстными векторами (*Cos*); г) возможность разрешения неоднозначности на основе лексических маркеров значения слова в контексте либо на основе лексико-семантических тегов его контекстного окружения. В ходе экспериментов нашла подтверждение гипотеза о большей эффективности разрешения неоднозначности с опорой на лексико-семантическую разметку корпуса текстов.

При развитии исследования был усовершенствован компьютерный инструмент автоматического разрешения лексико-семантической неоднозначности слов в контекстах, благодаря чему оказалось возможным проведение экспериментов а) с использованием не только лексической и семантической, но также и грамматической информации, извлекаемой из контекстов, определение наличия или отсутствия зависимости между данными критериями; б) с изменением ширины контекстного окна и с учётом границ синтагм (ранее на размер и наполнение контекста не накладывалось дополнительных ограничений); в) с установлением объёмов эталонных выборок пропорционально долям контекстов для разных значений в экспериментальной выборке (в предыдущих экспериментах объём эталонных выборок устанавливался более жёстко). Обсуждаемые эксперименты направлены на оценку эффективности разрешения неоднозначности с более гибкими параметрами и в условиях недостаточности исходных данных (например, в тех случаях, когда объём эталонного класса существенно меньше 500 контекстов – см. выше).

2. Лингвистические данные

Эксперименты по разрешению лексико-семантической неоднозначности проводятся на материале Национального корпуса русского языка (НКРЯ)[1]. В качестве тестовых лексем выбраны

[1] Публикации НКРЯ: **http://www.ruscorpora.ru/corpora-biblio.html**

имена существительные: *дом, орган, лук, глава, площадь, проспект, клетка, ключ, коса* и пр. Известна филиация значений данных слов, фиксируемая в лексико-семантической аннотации НКРЯ. При описании значений анализируемых лексем использовалась структура значений слов в ТСРЯ[1]. Каждому значению соответствует особая комбинация тегов, принятых в системе разметки НКРЯ[2]. Для рассматриваемых слов были сформированы выборки контекстов, присутствующих в НКРЯ.

Процедура анализа иллюстрируется на примере существительного *вид*, характеризующегося достаточно высокой частотностью и многообразием значений, среди которых есть абстрактные и конкретные; выявляемые на основе лексической, семантической и грамматической сочетаемости слова в контексте (см. табл. 1).

Таблица 1. Филиация значений слова *вид*

Значения	**Лексико-семантическая аннотация**	**Примеры**	**Число контекстов (∑=2866)**
m1	r:abstr t:perc der:v	***Вид*** *на озеро*	1144
m3	r:concr t:workart	*Альбом с* ***видами*** *Кавказа*	10
m5	r:abstr t:ment	***Виды*** *на урожай*	10
m11	r:abstr der:shift	*Подсолнухи в* ***виде*** *букета*	1075
m12	r:concr t:doc	***Вид*** *на жительство*	7
m21	r:abstr r:concr pt:set sc:X	*Отряды и* ***виды*** *животных*	617
m22	r:abstr	***Вид*** *глагола*	3

Эксперименты по разрешению неоднозначности слова *вид* проводились только для значений *m1*, *m11*, *m21*, представленных достаточным количеством контекстов (так, из рассмотрения были исключены значения *m3*, *m5*, *m12*, *m22*).

[1] ТСРЯ – Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. М., 1992.

[2] Система тегов: **http://www.ruscorpora.ru/corpora-sem.html**

3. Компьютерное обеспечение экспериментов

В экспериментах использовался компьютерный инструмент автоматической классификации лексики, адаптированный для разрешения неоднозначности слов в контексте. Проводится автоматическая классификация контекстов употребления слов в разных значениях с использованием векторной модели экспериментальной выборки. Реализован алгоритм классификации с учителем. Программное обеспечение разработано П.В. Паничевой на языке Python. В ходе работы программы производятся следующие процедуры: предобработка; машинное обучение; распознавание образов.

На этапе предобработки в экспериментальной выборке определяется число контекстов на каждое из значений слова. Для каждого из значений формируются эталонная выборка (случайным образом отобранные контексты со снятой неоднозначностью, где реализуется рассматриваемое значение) и тестовая выборка (контексты, для которых проводится автоматическое разрешение неоднозначности без учёта априорной лингвистической информации).

На этапе машинного обучения проводится формирование статистических образов для значений слова. Образ значения есть вектор в векторном пространстве, координаты которого определяются частотами встречаемости лексических маркеров значения, лексико-семантических или морфологических тегов контекстных элементов в эталонной выборке. Устанавливаются дистрибуции лексических маркеров, лексико-семантических и морфологических тегов в выборке.

На этапе распознавания образов тестовые контексты представляются как вектора в векторном пространстве. Измеряется расстояние между контекстными векторами и каждым из образов значений. Выбирается образ, к которому контекстный вектор расположен ближе всего. Анализируемому слову в контексте приписывается значение ближайшего образа.

В завершение проводится проверка качества распознавания: сравниваются результаты автоматической и ручной обработки контекстов, вычисляется доля правильных и ошибочных решений для каждого из значений.

4. Ход экспериментов

Были произведены эксперименты трёх типов, направленные на автоматическое разрешение неоднозначности с использованием (1) лексического (*lex*), (2) семантического (*sem*) и (3) грамматического (*gram*) критериев. Лексический критерий предполагает разрешение неоднозначности на основе лексических маркеров значений слов в контекстах (тег леммы), семантический критерий – разрешение неоднозначности на основе лексико-семантической разметки контекстов (теги первого значения слова), грамматический критерий – разрешение неоднозначности на основе морфологической разметки контекстов (грамматические теги). В ходе исследования необходимо было установить наличие или отсутствие зависимости между данными критериями.

В каждой из серий экспериментов происходило изменение параметров разрешения неоднозначности: а) объём эталонных выборок изменялся пропорционально общему числу контекстов для каждого из рассматриваемых значений (*A*=10%, *B*=15%, *C*=20%), объем тестовых выборок составил 20 контекстов; б) изменялась ширина контекстного окна [–*i*, +*k*], где $1 \le i, k \le 5$, $i = k \vee i \ne k$ (допускается как симметричное, так и асимметричное окно); в) обработка контекстов проводилась с учётом границ синтагм: предварительные тесты показали, что в данном режиме качество распознавания значений возрастает на 0,5%…1% по сравнению с обычным режимом анализа контекстов (без учёта знаков препинания).

Во всех экспериментах близость контекстных векторов по отношению к образам определялась с помощью меры *Cos* как наиболее надёжной по сравнению с мерами Евклида и Хемминга.

5. Результаты экспериментов

5.1. Оценка точности результатов при изменении параметров экспериментов

Значения точности *P* определялись как отношение объёма тестовых выборок для каждого из значений к числу контекстов, по которым были приняты верные решения об их принадлежности к тому или иному образу. Также вычислены значения $P_{ср}$ в сериях экспериментов. Результаты приведены в табл. 2.

Таблица 2. Оценка точности результатов: *вид*

P	***lex***			***sem***			***gram***		
	m1	***m11***	***m21***	***m1***	***m11***	***m21***	***m1***	***m11***	***m21***
A	0,56	0,82	0,61	0,61	0,32	0,66	0,65	0,74	0,7
B	0,67	0,69	0,52	0,57	0,68	0,72	0,69	0,68	0,77
C	0,65	0,76	0,59	0,56	0,51	0,71	0,6	0,83	0,72
$P_{ср}$	***lex***			***sem***			***gram***		
A	0,66			0,53			0,7		
B	0,63			0,66			0,71		
C	0,67			0,59			0,72		

Наибольшая точность результатов разрешения неоднозначности достигается в экспериментах с использованием грамматического критерия ($P_{ср}$=0,7…0,72), на втором месте – лексический критерий ($P_{ср}$=0,63…0,67), на третьем – семантический критерий ($P_{ср}$=0,53…0,66). При возрастании объёма эталонных выборок наблюдается весомое увеличение точности.

Замечено, что значение *m1* при любых параметрах экспериментов распознаётся хуже, чем значения *m11* и *m21*. Пара значений *m11* и *m21* дифференцируется по типу: при использовании лексического критерия оказывается выше точность распознавания значения *m11*, а при использовании семантического критерия возрастает точность распознавания значения *m21*. Это подтвер-

ждает гипотезу о специализации критериев разрешения неоднозначности с точки зрения типов лексических значений.

Существенное влияние на точность результатов разрешения неоднозначности оказывает ширина контекстного окна. Пока не удалось установить наилучшее соотношение $[-i, +k]$, однако эксперименты позволяют высказать предположение, что вне зависимости от выбранного критерия точность распознавания значений выше при асимметричном контекстном окне, где $i \leq 2$, $2 \leq k \leq 4$. В отдельных случаях точность распознавания возрастает до 0,95…1.

5.2. Оценка полноты результатов при изменении параметров экспериментов

Значения полноты R определялись как отношение объёма тестовых выборок для каждого из значений к числу контекстов, по которым были приняты верные и ошибочные решения об их принадлежности к тому или иному образу. Также вычислены значения $R_{ср}$ в сериях экспериментов. Результаты приведены в табл. 3.

*Таблица 3.*Оценка полноты результатов: *вид*

R	***lex***			***sem***			***gram***		
	m1	***m11***	***m21***	***m1***	***m11***	***m21***	***m1***	***m11***	***m21***
A	0,86	0,99	0,95	0,93	0,99	1	1	1	1
B	0,88	0,97	0,97	0,96	0,99	1	1	1	1
C	0,86	0,99	0,93	0,93	0,99	1	1	1	1
$\boldsymbol{R_{ср}}$	***lex***			***sem***			***gram***		
A	0,93			0,97			1		
B	0,94			0,99			1		
C	0,94			0,97			1		

Можно заметить, что наибольшую полноту обеспечивает грамматический критерий ($R_{ср}$=1), второе место занимает семантический критерий ($R_{ср}$=0,97…0,99), на третьем месте – лексичес-

кий критерий ($R_{ср}$=0,93…0,94). При возрастании объёма обучающих выборок наблюдается незначительное увеличение полноты.

Зарегистрировано снижение полноты в распознавании значения *m1* как по лексическому (*R*=0,86…0,88), так и по семантическому критериям; значение *m21* распознаётся по семантическому критерию (*R*=1) с большей полнотой, чем по лексическому (*R*=0,93…0,97); значение *m11* распознаётся по лексическому и семантическому критериям примерно одинаково (*R*=0,97…0,99). Замечено, что наибольшее снижение полноты происходит при экспериментах с ограниченным левым контекстом (контекстное окно [–1, +*k*], где *k*=1…5).

5.3. Анализ сложных случаев

Ошибочные решения, потери данных и отсутствие решений о принадлежности контекстов к тому или иному образу объясняются следующими причинами:

а) недостаточность лексических, семантических и/или грамматических признаков в контексте для правильного определения значения:

– исходное значение *m1*, распознанное значение *m11*, критерий *sem*: [419] *Клумбы опустели и имели беспорядочный* ***вид****.*

– исходное значение *m11*, распознанное значение *m1*, критерий *sem*: [1408] *Гидротальки́т встречается как в* ***виде*** *отдельных идиоморфных кристаллов размером до 5 миллиметров в поперечнике, так и в сростках и в друзоподобных агрегатах.*

– исходное значение *m21*, распознанное значение *m11*, критерии *lex*, *gram*: [2341] *Например, в ФРГ федеральные структуры исполнительной власти разделены на три* ***вида****: министерства, ведомства и службы*.

– исходное значение *m1*, распознанное значение 0, критерий *lex*: [134] ***Вид*** *Глюкала поразил пришедших.*

– исходное значение *m11*, распознанное значение 0, критерий *sem*: [2241] *Он всё и произнес лишь для того, чтобы доставить маленькому человеку страдания в самом невыносимом* ***виде****.*

– исходное значение *m21*, распознанное значение 0, критерий *sem*: [2916] *Существовал, однако,* ***вид*** *крамолы неистребимой.*

б) употребление анализируемого слова в составе устойчивого сочетания или конструкции, например:

– исходное значение *m1*, распознанное значение 0 или *m21*, критерии *lex*, *sem, gram*: [1121] *Порой Елене казалось, что все явления и все предметы можно описать в трёх позициях: анфас, профиль,* ***вид*** *сверху.*

– исходное значение *m11*, распознанное значение *m1*, критерии *lex*, *gram*: [1686] *Обычно они существуют / плохие или хорошие / в том или ином виде / но существуют / установленные / может быть / не совсем верно с точки зрения иерархии законодательных актов.*

– исходное значение *m1*, распознанное значение 0, критерий *lex*: [96] *Однако* ***виду*** *никогда не подаст.*

– исходное значение *m11*, распознанное значение 0, критерий *sem*: [1798] *Герасим Николаевич говорит: «Доктор, я не баба, видел* ***виды****… говорите, она?»*

– исходное значение *m21*, распознанное значение 0, критерий *lex*: [2573] *Сами-то того не знаете* ***вида****…*

5.4. Оценка корреляции параметров экспериментов

Был проведён корреляционный анализ экспериментальных данных, направленный на определение зависимости между (1) лексическим (*lex*), (2) семантическим (*sem*) и (3) грамматическим (*gram*) критериями разрешения неоднозначности с учётом изменения объёмов эталонных выборок (*A*=10%, *B*=15%, *C*=20%). Расчёты осуществлены в вычислительной среде *Mathcad*. Определены значения коэффициента корреляции Пирсона *Corr* (*X*, *Y*). Результаты представлены в табл. 4.

Таблица 4. Результаты корреляционного анализа: *вид*

Corr (*A*1, *A*2) = –0,704	*Corr* (*B*1, *B*2) = –0,045	*Corr* (*C*1, *C*2) = –0,309
Corr (*A*1, *A*3) = –0,011	*Corr* (*B*1, *B*3) = –0,148	*Corr* (*C*1, *C*3) = 0,364
Corr (*A*2, *A*3) = 0,175	*Corr* (*B*2, *B*3) = 0,377	*Corr* (*C*2, *C*3) = –0,109

Данные свидетельствуют об отсутствии устойчивой зависимости между лексическим, семантическим и грамматическим критериями. Исключение составляет достаточно сильная обратная зависимость между лексическим и семантическим крите-

риями при объёме эталонных выборок 10%: *Corr* = –0,704, однако в остальных случаях прямая или обратная зависимость не превышает значения |*Corr*| = 0,377. Это означает, что можно ожидать повышения качества разрешения неоднозначности в экспериментах с комбинированными критериями (*lex*+*sem*, *sem*+*gram*, *lex*+*gram*, *lex*+*sem*+*gram*).

6. Выводы

Эксперименты по статистическому разрешению лексико-семантической неоднозначности русских имён существительных в корпусе с гибкими условиями привели к следующим результатам:

1) произведен сравнительный анализ эффективности лексического, семантического и грамматического критериев разрешения неоднозначности; подтверждена гипотеза о специализации данных критериев в отношении значений различных типов; показана несколько большая надёжность грамматического критерия по сравнению с лексическим и семантическим; установлено отсутствие устойчивых связей между тремя критериями;

2) установлено повышение эффективности разрешения неоднозначности при формировании эталонных выборок пропорционально объёму экспериментальных выборок; подтверждено улучшение результатов распознавания значений при анализе контекстов с учётом границ синтагм; показана зависимость точности и полноты результатов разрешения неоднозначности от выбора критериев разрешения неоднозначности, объёма эталонных классов и ширины контекстного окна; высказано предположение о приемлемом размере контекстного окна, обеспечивающем достаточную точность результатов с незначительной потерей полноты;

3) подтверждена возможность эффективного разрешения неоднозначности при относительно малых эталонных классах; описаны сложные случаи ошибочных решений и потери данных.