

Fuchs-Gambeck, M. (2010). *Rammstein. Goryashchiye serdtsa* [Rammstein. Burning heart] Moscow: AST.

УДК 81.33

<https://doi.org/10.25076/vpl.36.03>

А.А. Гаджиев

А. К. Хмелёв

Московский государственный областной университет

## АЛГОРИТМ ЛЕСКА И СИСТЕМА BABELFY ДЛЯ ДИЗАМБИГУАЦИИ

Дизамбигуация является актуальным направлением исследований в сфере теоретической, прикладной и компьютерной лингвистики. В настоящее время задача качественного снятия лексической многозначности не решена, тем не менее, существует ряд подходов к дизамбигуации. В статье описан эксперимент по анализу работы систем разрешения лексической многозначности – алгоритма Леска и системы Babelfy. Системы, выбранные для работы, основаны на разных подходах к дизамбигуации. Алгоритм Леска работает на пакете библиотек и программ NLTK, Babelfy - на основе семантической сети Babelnet. Тестирование проводилось с использованием собранной выборки предложений, содержащих многозначные слова, фразовые глаголы, омонимы, другие неоднозначные лексические единицы. В ходе исследования проводился анализ качества работы систем, для каждой из них представлен коэффициент эффективности. В соответствии с проведенным статистическим анализом ошибок можно сделать вывод о недостаточно высоком качестве работы систем снятия многозначности. В заключении описаны возможные причины ошибок систем дизамбигуации и предложено решение по их улучшению.

Ключевые слова: обработка текста, семантическая сеть, дизамбигуация, значение, контекст, синсет, тезаурус.

UDC 81.33  
<https://doi.org/10.25076/vpl.36.03>  
A.A. Gadzhiev  
A.K. Khmelev  
Moscow Region State University

## LESK ALGORITHM AND BABELFY SYSTEM FOR DISAMBIGUATION

*Disambiguation is a relevant scientific field of research in language theory and natural language processing. Nowadays the task of qualitative removing of ambiguity is still not solved, nevertheless, several approaches to word sense disambiguation are available. The article describes the experiment of testing systems of word sense disambiguation – the Lesk algorithm and Babelfy system. The systems belong to different approaches. The Lesk algorithm runs on the NLTK library and software package and Babelfy is based on the Babelnet semantic network. The testing was conducted using several sentences containing ambiguous words, expressions, phrasal verbs, homonyms and other ambiguous constructions. During the experiment, the quality of the work of the systems was evaluated. According to the statistical analysis of errors it can be concluded, that the quality of work of systems for removing ambiguity is not high enough. In conclusion, the possible causes of errors of disambiguation systems are described and a solution to improve them is proposed.*

*Key words:* text processing, semantic network, disambiguation, meaning, context, synset, thesaurus.

### **Введение**

Многозначность как возможность слова иметь множество значений, с одной стороны, отражает богатство языка, но, с другой, является препятствием к его пониманию, создает трудности для всех сфер направления обработки текстов. Направление, занимающееся решением этой проблемы, получило название Word sense disambiguation или разрешение лексической многозначности, суть которого состоит в установлении значения слов в соответствии с контекстом.

Решение задачи может быть сведено к 2-м пунктам:

1. определить, какие значения может иметь каждое слово, относящееся к тексту;

2. выбрать самое подходящее значение, исходя из контекста.

Развитие области в настоящее время замедляется, из-за отсутствия революционных улучшений и трудностей внедрения существующих систем на нижние уровни обработки естественного языка (downstream NLP applications) (Iacobacci, Navigli, Pilehvar, 2016). Также само по себе выявление реальных улучшений по сравнению с существующими подходами является трудной задачей из-за отсутствия единой системы оценивания, что препятствует проведению прямых и справедливых сопоставлений между системами (Camacho-Collados, Navigli, Raganato, 2017). Несмотря на данные препятствия, исследование разрешения многозначности продолжается, совершенствуются старые и появляются новые методы и основанные на них системы дизамбигуации. Основные используемые методы – с учителем, без учителя, метод, основанный на знаниях, и гибридный (Butnaru, Ionescu, & Shotgun, 2019). В данной работе проводилось тестирование систем разного типа, и главной задачей было оценить способность систем эффективно выполнять задачу снятия многозначности.

### **Теоретические основания**

Компьютерная лингвистика как самостоятельное научное направление, как и большинство других наук, основанных на использовании информационных технологий, возникла и оформилась в середине XX века в связи с появлением ЭВМ и развитием кибернетики (Navigli, 2009).

Позднее, благодаря успеху Джорджа Таунского эксперимента (1954), множество лингвистов заинтересовалось проблемой машинного перевода, а также проблемой анализа и обработки текстов на естественном языке, изучением общих законов естественного языка — его структуры и функционирования. Одним из направлений внутри обширной проблемы обработки текстов является разрешение лексической многозначности или, по-другому, Word sense disambiguation.

Стоит упомянуть, что еще до Джорджа Таунского эксперимента Уивер в своём манифесте The „Translation“ memorandum определил многозначность как одну из 4 проблем машинного

перевода (Weaver, 1949, с.2). Он также предполагал, что проблема может быть решена при помощи изучения непосредственного контекста многозначного слова. Он проиллюстрировал своё предложение примером:

«Если рассматривать слова в книге по одному через непрозрачную маску с отверстием шириной в одно слово, то, очевидно, невозможно определить значения слов. “Fast” может означать “быстро” (rapid), а может означать “неподвижно” (motionless), и невозможно определить, что именно.

Но если удлинить щель в непрозрачной маске до тех пор, пока будет видно не только центральное слово, но и N слов с обеих сторон, то, если N достаточно большое, можно однозначно определить значение» (Weaver, 1949, там же).

Проблема заключалась в определении длины необходимого контекста, Уивер ожидал, что это будет варьироваться от одного предмета к другому. Он считал, что относительно небольшое количество существительных, глаголов и прилагательных на самом деле неоднозначны. Тем не менее, эта позиция не подтвердилась при анализе естественных языков (Booth & Locke, 1955).

Уивер и другие исследователи понимали степень важности и сложности проблемы многозначности. Во многом, из-за её влияния на области применения, ведь WSD ещё с давних пор задумывалась как задача, решение которой - главное препятствие на пути к почти идеальному машинному переводу.

Слова в развитых языках могут быть как однозначными, так и многозначными. Однозначность или моносемия (от греч. *monos* — один + *sema* — знак), как писал Розенталь, это наличие у слова только одного значения (Розенталь, Теленкова, 1985).

Многозначность или полисемия (но не во всех случаях??) — наличие у слова больше одного, нескольких значений.

Турдаков, автор диссертации о WSD, приводит пример, что из 121 самых употребляемых имен существительных английского языка каждое имеет в среднем около 7.8 значений, представленных в тезаурусе WordNet (Турдаков, 2010).

Существование многозначности обусловлено историческим развитием языка, когда развитие лексики происходит не путём образования новых слов, а с помощью появления новых значений

у существующих. Вдобавок к этому, многозначность происходит из способности слова употребляться в переносных значениях. Может происходить перенос номинации с одного предмета на другой, при наличии у предметов общих признаков. Отличие многозначных слов от омонимов, слов, совпадающих по звучанию, но различающихся по значению слов, заключается в присутствии семантической связи между значениями, что позволяет причислять их к значениям одного слова.

Трудно представить многозначность как простое и однородное явление. Принято выделять несколько типов полисемии, для работы с каждым из которых используются различные методы:

Морфологическая (грамматическая) многозначность. Подразумевается отнесение значений слова к разным частям речи, к примеру, англ. “face” как существительное «лицо» и как глагол «сталкиваться». Морфологическая многозначность - главный объект исследования задачи частеречного определения (part of speech tagging, POV).

Синтаксическая многозначность - возможность описания предложения при помощи двух и более синтаксических структур. Пример: “burning firework can be unsafe” («вспыхнувший фейерверк может быть небезопасным» или «поджигать фейерверк может быть небезопасно»).

Прагматическая неоднозначность возникает, когда используются местоимения или специальные существительные вроде one, another (еще один). Например, в предложении «Он уронил гитару на ногу и сломал её» нельзя однозначно сказать, что именно было сломано – гитара или нога, но при этом в обоих случаях присутствует одинаковая синтаксическая структура. В данной работе синтаксическая и прагматическая многозначности не рассматриваются, так как в ней не затрагивается уровень синтаксического анализа.

Лексическая многозначность. Классическое определение – это фундаментальное свойство естественных языков, заключающееся в том, что каждое слово может иметь более одного значения. Именно разрешение лексической многозначности является основным вариантом перевода английского термина Word sense disambiguation. Также может употребляться аббревиатура WSD или заимствование “дизамбигуация”. Именно этот тип

многозначности будет рассмотрен в работе дальше.

Иногда в качестве отдельного типа выделяется семантическая многозначность, как возможность употребления слова в переносном значении (пример: «зверь» – жестокий, свирепый человек). Но для её разрешения, в большинстве случаев, могут применяться методы разрешения лексической многозначности.

В своей диссертации Турдаков сводит решение задачи разрешения лексической многозначности к 2-м необходимым пунктам (Турдаков, 2010):

1. определение возможных значений;
2. выбор подходящего значения, исходя из контекста.

Также он указывает на то, что для формализации задачи требуется дать дефиниции «значения» и «контекста», потому что их отсутствие создаёт дополнительные сложности (Розенталь, Теленкова, 1985).

Несмотря на сказанное, не представляется возможным ограничиться одним строгим определением “значения” или “контекста”. Можно рассмотреть существующие подходы и точки зрения, касательно этих понятий:

Способа точно определить, где заканчивается одно значение и начинается другое, не существует. Обычно при таком вопросе опираются на словари и другие списки слов, но данные в разных источниках могут различаться. Поэтому существует другой способ, при котором анализируются способы употребления слов, собранных в лингвистический корпус, и значения описываются на основе этого анализа. В настоящее время над текстовыми корпусами языков ведётся активная работа, в их пополнении нередко участвуют студенты лингвистических факультетов.

На информации, которую предоставляет контекст слова, основывается выбор значения этого слова. В работах по дизамбигуации контекст может рассматриваться как, в целом, окружение слова, так и с учётом некоторых отношений элементов этого окружения с целевым словом, такие как расстояние до цели, синтаксические связи, орфографические свойства, семантические категории. Различают микро-(несколько слов) и макроконтекст (несколько предложений, текст) (Розенталь, Теленкова, 1985). Также может учитываться контекст, определяемый областью знаний, для которой решается задача снятия многозначности.

Чаще всего используется микроконтекст, однако, вопрос каким должен быть его размер, остается спорным.

Роберто Навигли пишет в своей работе (Navigli, 2010), что, если пренебречь пунктуацией, то можно рассматривать текст  $T$  как последовательность слов ( $w_1, w_2, \dots, w_n$ ), и дизамбигуацию можно описать как задачу присвоения соответствующего смысла (или смыслов) всем или некоторым словам в  $T$ .

Проблема дизамбигуации, которая является частью проблемы обработки естественного языка, в целом, и по сей день имеет статус неразрешенной. Направление развивалось до некоторого предела, и показатели успешности его исследований можно охарактеризовать как “сравнительно-эффективные”. По достижению упомянутого предела, прогресс в исследованиях сильно замедлился. Причиной этого во многом являются неразрешенные проблемы, которые связаны, в первую очередь, с языковыми особенностями человеческой речи, и другими проблемами.

Научные исследования, посвященные разрешению лексической многозначности, входят в спектр интересов прикладной и компьютерной лингвистики. Прикладная лингвистика - область, развивающая и использующая технологии для разрешения многозначности и обработки речи и текста, в целом, а компьютерная лингвистика ставит своей целью использование информационных технологий для усовершенствования и движения вперед лингвистической теории (Турдаков, 2010).

Также дизамбигуация входит в перечень многочисленных проблем, связанных с исследованием искусственного интеллекта (AI). Нужно отметить, что со сферой искусственного интеллекта дизамбигуация соотносится не только по близости сферы работы, но и по уровню сложности задачи (Navigli, 2010).

Актуальность той или иной проблемы, задачи коррелирует с возможностью её применения в разных областях, количеством этих областей и тем, что из себя представляют эти области. Дизамбигуация применяется во многих приложениях для обработки естественного языка, в частности:

- Машинный перевод. Осуществление правильного перевода слов во многом зависит от понимания их смысла. Соответственно,

отсутствиедельно работающих механизмов распознавания смыслов приводит к неправильному переводу.

- Информационный поиск. Использование дизамбигуации необходимо для увеличения релевантности результатов запросов в поисковых системах путём исключения из рассмотрения результатов употребления какого-либо из слов запроса в неинтересующем в текущий момент значении.
- Контент-анализ. Другими словами, анализ распределения категорий слов в текстовых коллекциях - слов по определенной теме. Для создания таких категорий требуется установить конкретные значения слов.
- Обработка текстов и речи. Дизамбигуация повышает точность анализа текстов, их классификации и кластеризации, способствует исправлению ошибок; оказывает помощь в правильном воспроизведении слов при синтезе текстов, распознавании речи.

Успехи в решении этих задач или движение в сторону их оптимизированной работы происходят только до определенной степени эффективности, что связано с проблемами, затрудняющими разрешение многозначности.

Возникающие сложности дизамбигуации чаще всего связаны с особенностями речи и психологии человека (Navigli, 2010):

- Составление словарей. Словари различаются, в зависимости от составителей, а также других обстоятельств, одно из которых - вопрос "что считать за значение слова?". Бывает трудно определить, когда слова относятся к одному значению или представляют разные, что приводит к несовпадающему разделению на смыслы в разных словарях и тезаурусах.
- Проблема определения частей речи (POV). При любом реальном тестировании, частеречная разметка и разметка значений очень тесно связаны, чем потенциально создают друг для друга ограничения. До сих пор ведутся споры о том, следует ли разделять эти две задачи или они требуют совместного решения. В последнее время эти проблемы исследуются отдельно. Хотя обе задачи используют и дизамбигуацию, и частеречную разметку, алгоритмы одной неприменимы в должной мере к другой. Во многом из-за того, что для определения части речи слова

требуется от 1-го до 3-х соседних слов, тогда как для разрешения многозначности может потребоваться гораздо больше.

- Человеческий фактор, согласование ручных результатов. Результаты тестов систем дизамбигуации обычно сопоставлялись с результатами работы человека. Но разметка смыслов является труднодостижимой для человека в полной мере, ведь запомнить все смыслы слов просто невозможно. Также в такой разметке у всех людей будут разные результаты.

- Здравый смысл. Специалист по искусственному интеллекту Дуглас Ленат в своем докладе “Computers versus common sense” в рамках образовательной программы “GoogleTechTalks” говорит о важности здравого смысла при обработке текстов (Lenat). В пример можно привести следующие предложения:

«Naruto and Sasuke are brothers.» — (они являются братьями по отношению друг к другу).

«Naruto and Sasuke are fathers.» — (каждый независимо является отцом).

Для анализа таких значений необходимы подобные знания о мире и обществе.

- Зависимость от поставленной задачи. Существуют случаи, в которых многозначность слова никак не влияет на его перевод. (Пример: mouse=животное/устройство) Для разных случаев требуются разные алгоритмы, и эффективно работающий в одной ситуации алгоритм в другой будет просто бесполезен.

Существует несколько основных методов, над которыми продолжается работа.

Методы, основанные на знаниях. Эти методы преимущественно полагаются не на корпуса текстов, а на словари, тезаурусы, лексикографические базы данных. Самыми известными и крупными базами являются английская семантическая сеть Wordnet - базовыми языковыми единицами в ней являются синонимические ряды - синсеты, которые объединяют слова со схожими значениями и являются узлами сети.

К этому методу относится алгоритм Леска и его усовершенствованные версии. Суть работы метода в сопоставлении слова и его смысла в словаре путём поиска значения слова в списке словарных определений с учетом

контекста, где слово использовалось. Алгоритм основан на утверждении, что многозначное слово и его окружение относятся к одной теме (Lesk, 1986).

Методы, основанные на знаниях, легко адаптируются к документам из любых источников и могут использоваться с разными языками.

Методы полного и частичного обучения с учителем использует техники машинного обучения для составления классификатора из вручную аннотированных по смыслу наборов данных. Классификатор работает с одним словом и присваивает соответствующий смысл каждому экземпляру этого слова. Тестовый набор для изучения классификатора снабжен базами примеров, в которых заданное целевое слово вручную помечается значением из словаря. Метод основывается на предположении, что контекст целевого слова предоставляет достаточно информации, чтобы определить конкретное значение его употребления.

Недостатком метода является ручная разметка текстов для обучения. Так однажды был произведен расчёт, что для получения высокоточной системы широкого охвата неоднозначности нам, вероятно, нужен корпус из примерно 3,2 миллиона слов с меткой смысла. Человеческие усилия по созданию такого учебного корпуса могут быть оценены в 27 человеческих лет, при пропускной способности одного слова в минуту.

Методы обучения без учителя имеют потенциал разрешить проблемы методов, основанных на масштабных, вручную аннотированных ресурсах со значениями слов. Опираются на утверждения, что «схожие значения встречаются в схожих контекстах» и могут быть извлечены из текста при помощи кластеризации, основываясь на некоторой мере схожести контекстов (Schütze, 1998). Недостаток таких систем в том, что они не могут полагаться на общие базы значений, так как не используют словари. Однако на эти системы возлагают большие надежды.

В настоящее время большое внимание уделяется методам, использующим в качестве корпуса интернет пространство, по причине того, что такой корпус имеет очень большой размер и является мультилингвистическим (Navigli, 2010).

В пример можно привести диссертацию Турдакова,

содержащую методы дизамбигуации на основе сетей документов, таких как Википедия (Турдаков, 2010).

Также нельзя обойти стороной новую систему дизамбигуации Comprehendo, от компании Babelscape, демо-версия которой, по информации на официальном сайте, должна быть скоро доступна. Система базируется на мультилингвистическом графе знаний World Atlas и семантической сети Babelnet, которая сейчас работает в live-режиме и обновляется регулярно.

Возможно, именно эта система сможет справиться с главным препятствием в разрешении лексической многозначности - недостатком знания.

### **Эксперимент**

В работе проводилось тестирование двух систем: одного из самых первых алгоритмов разрешения лексической многозначности – алгоритма Леска и более новой системы Babelfy.

Алгоритм Леска возник в 1986 году, суть метода состоит в сопоставлении слова и его смысла в словаре путём поиска значения в словаре с учетом контекста. (Lesk, 1986).

Была установлена версия алгоритма Simple Lesk и приложенный к нему пакет данных для использования в Python с веб-сервиса GitHub. Он работает на пакете программ и библиотек для обработки естественного языка NLTK. В функции этой версии входит решение следующих задач: установление синсета – синонимического ряда, к которому принадлежит слово, вывод его определения из словаря, также установление гипонимов и гиперонимов и дизамбигуация всех сущностей в предложении (Ayetiran, Agbele, 2018). Никаких ресурсных дополнений для этого алгоритма не требует (Алексеевский, 2018).

Руководствуясь инструкцией по сборке, с веб-сервиса GitHub, была установлена библиотека NLTK и пакет “pywsd”, содержащий модуль simple\_lesk. Достаточно ввести команду from pywsd.lesk import simple\_lesk, чтобы модуль simple\_lesk был импортирован в Python и алгоритм можно было применять. Его работа осуществляется следующим образом: с помощью команды sent = " в систему вводится фраза, предложение или несколько предложений. Далее командой ambiguous = " маркируется объект дизамбигуации - слово или выражение, неоднозначность которого требуется разрешить. Нужно также указать часть речи слова или,

если слов несколько, то часть речи главного из них, прописав команду `answer = simple_lesk(sent, ambiguous, pos="")`. После чего с помощью `print answer` в следующей строке можно получить ответ в виде синсета, к которому относится объект и, впоследствии, введя команду `print answer.definition()` получить развернутое определение данного объекта - разрешенную неоднозначность.

Алгоритм хорошо размечает значения в известных конструкциях в контекстах с использованием часто употребляемых с ним слов, как на рисунке 1. В предложениях по типу примера 1:

1) 'I went to the bank to deposit my money'  
выбирается правильный синсет  
'depository\_financial\_institution.n.01' и верное значение 'a financial institution that accepts deposits and channels the money into lending activities'.

Также правильно разрешается многозначность с разными частями речи, например, "fire=огонь/увольнять". Алгоритм привязан к точности формулировок, и как только в предложении выше фраза "to deposit my money" была заменена на менее официальную "to get cash", программа ошиблась.

Также программу может запутать изменение целевого слова. Например, слово "cash", выбранное целевым в том же предложении без каких-либо изменений, принимается за американского кантри-певца, как в примере 2:

```
2) >>> print(answer)
Synset ('cash.n.03')
>>> print(answer.definition)
United States country music singer and songwriter (1932-2003)
```

Программа правильно воспринимает известные употребляемые фразовые глаголы, такие как "look forward to", "wash away", "strike down", но не определяет идиомы, наподобие "be a peeping Tom".

При этом программа распознаёт некоторые фразы из художественного текста в переносном значении, метафоры, такие как "wash away from the soul".

Также учитывается контекст величиной больше одного предложения. Для примера возьмём предложение "I like covering my nails with bright red.", где слово "nail" употребляется в значении "ноготь", и предложение "He hammered the last nail into the coffin.",

где оно означает “гвоздь”. Составим текст из 2-х предложений – “I like covering my nails with bright red. And then I hammer them into the coffin.” Система верно определит, что речь в данном тексте шла о гвоздях, а не ногтях.

Для того, чтобы произвести дизамбигуацию целого предложения, то есть его каждого отдельно взятого слова, требуется другой набор действий. С помощью from pywsd import disambiguate из того же пакета pywsd импортируется модуль disambiguate. Следующий шаг — это, непосредственно дизамбигуация. Прописывается команда disambiguate (""), с помощью которой в систему вводится предложение, все члены которого дизамбигуируются. На выходе получается ответ в виде каждого слова и синсета, к которому оно относится последовательно через запятую. В случае если объект не является неоднозначным, не входит ни в один из синсетов, то выводится ответ “None”. Смотрите примеры 3, 4 и 5:

3) >>> from pywsd import disambiguate

>>> disambiguate('He hammered the last nail into the coffin.')

[('He', None), ('hammered', Synset('hammer.v.01')), ('the', None), ('last', Synset('last.s.09')), ('nail', Synset('nail.n.02')), ('into', None), ('the', None), ('coffin', Synset('coffin.n.01')), ('.', None)]

4) >>> disambiguate('During the storm, the ship found safety in a quiet harbor behind the rocks. ')

[('During', None), ('the', None), ('storm', Synset('storm.n.03')), ('.', None), ('the', None), ('ship', Synset('ship.n.01')), ('found', Synset('witness.v.02')), ('safety', Synset('safety.n.06')), ('in', None), ('a', None), ('quiet', Synset('quiet.s.03')), ('harbor', Synset('seaport.n.01')), ('behind', Synset('buttocks.n.01')), ('the', None), ('rocks', Synset('rock\_candy.n.01')), ('.', None)]

5) >>> disambiguate('I went to the bank to deposit my money')

[('I', None), ('went', Synset('run\_low.v.01')), ('to', None), ('the', None), ('bank', Synset('depository\_financial\_institution.n.01')), ('to', None), ('deposit', Synset('deposit.v.02')), ('my', None), ('money', Synset('money.n.03'))]

Как мы можем увидеть, на экран выводятся только синсеты, в

которые входят в дизамбигуируемые слова, а не определения их значений. Это расценивается нами, как недостаток, так как не предоставляется возможности оценить правильность результата до конца, ведь нельзя исключать вероятность, что в правильно выбранном синсете системой может быть выбрано неправильное значение.

Алгоритм исправно работает в стандартных случаях, но ошибается в менее употребляемых конструкциях. Причина таких ошибок – ограничение базы знаний и простота алгоритма.

Babelfy от компании Babelscape – это единая мультилингвистическая система, основанная на графах системы разрешения лексической многозначности и связывания сущностей (Entity Linking). Базируется на многоязычной семантической сети BabelNet 3.0. Система поддерживает 271 язык, включая и естественные (английский, русский, немецкий, французский, польский, шведский, китайский и т. д.) и искусственные (Эсперанто).

Хотя в основе системы Babelfy тоже заложен сложный алгоритм, она уникальна тем, что является понятной и возможной для использования рядовым пользователем. Она не требует ручного ввода данных, таких как ключевое слово и его часть речи в алгоритме Леска. Требуется только ввести предложение.

Система производит разметку сразу всех слов и не требует выбирать какое-то конкретное слово, как целевое, хорошо разрешает многозначность в некоторых устойчивых конструкциях, как в примере 6:

6) A bolt of lightning hit the wooden tower and totally burnt it.

Найден ные сущности	Значения
bolt of lightning	thunderbolt <i>A discharge of lightning accompanied by thunder</i> lightning <i>To produce lightning.</i> latch <i>Catch for fastening a door or gate; a bar that can be lowered or slid into a groove</i>
hit	<i>Cause to move by striking</i>

wooden tower	wooden tower <i>Tower made out of wood</i> wooden <i>Made or consisting of (entirely or in part) or employing wood</i> tower <i>A structure taller than its diameter; can stand alone or be attached to a larger building</i>
totally	<i>To a complete degree or to the full or entire extent ('whole' is often used informally for 'wholly')</i>
burnt	<i>Ruined by overcooking</i>

Таблица 1.

Слово bolt, основное значение которого “болт, винт”, употреблено в другом значении - “вспышка”. Программа определяет многозначное слово правильно, так как в предложении оно употреблено в словосочетании “a bolt of lightning”. Стоит отметить, что для некоторых сущностей предложены значения как для отдельных единиц, так и значения в составе словосочетаний, например, “wooden tower”.

Значение сущности “burnt” определено не верно, в примере Babelfy определила его как «испорченный чрезмерной температурной обработкой во время приготовления», что явно не подходит к описанию результатов удара молнии.

7) You can stay stuck in traffic jams in Moscow for ages.

Найден ные сущности	Значения
stay	rest <i>Stay the same; remain in a certain state</i>
stuck	wedge <i>Put, fix, force, or implant</i>
traffic jams	traffic jam <i>A number of vehicles blocking one another until they can scarcely move</i> traffic <i>The aggregation of things (pedestrians or vehicles)</i>

	<i>coming and going in a particular locality during a specified period of time</i>
Moscow	Moscow <i>A city of central European Russia; formerly capital of both the Soviet Union and Soviet Russia; since 1991 the capital of the Russian Federation</i>
ages	age <i>A prolonged period of time</i>

Таблица 2.

В примере 7 разрешение многозначности слова “jam” с основным значением “джем”, происходит по похожему принципу. Выбирается правильное значение “пробка”, так как слова стоят в словосочетании с прилагательным “traffic”.

Однако, может ошибиться в некоторых примитивных случаях, таких как в примере 8:

8) I went to the bank to deposit my money.

Определенное значение – “bank=банк/берег” – “Sloping land (especially the slope beside a body of water.” Очевидно, что речь в предложении шла о слове “bank” в значении “банк”, так как оно было употреблено в контексте со словосочетанием “to deposit money”.

Недостатком является то, что чаще выбираются либо наиболее общие значения, либо самые употребляемые. Программа также учитывает контекст более одного предложения, как и алгоритм Леска, но имеет больше возможностей в этом отношении. Алгоритм Леска просто определяет из широкого контекста целевое слово, а система Babelfy в состоянии работать с большим текстом и не акцентировать внимание на каком-то конкретном слове.

Система знает некоторые идиомы и фразовые глаголы, но при этом может не знать другие, не менее употребляемые. Так не была распознана конструкция “look forward to” в примере 9:

9)I'm looking forward to buy a pool table in my country house.

В этом предложении Babelfy не выделила “look forward to” как целое выражение, были представлены отдельные значения для слов “look” и “forward”. Для слова “forward” также был предложен диалектный вариант “forrad”. При этом для словосочетания

“country house” Babelfy выдала как общее значение “A house (usually large and impressive) on an estate in the country”, так и отдельные значения для слов=элементов словосочетания. “Pool table” было описано похожим образом, первым было предложено значение «бильярдный стол», затем- значения для отдельных компонентов, при этом, для слова “pool” это был не «бильярд», а «бассейн».

Но при этом была определена ранее упомянутая идиома “a peeping Tom”, правда в не совсем корректном значении, что показано в примере 10:

10) He was a peeping Tom for a group of people from his window.

Найден ные сущности	Значения
peeping Tom	<p>peeping Tom  <i>A viewer who enjoys seeing the sex acts or sex organs of others</i></p> <p>peep  <i>Look furtively</i></p> <p>Tom  <i>The tom is a plucked lamellophone used in the traditional music of the Nuer and Anuak ethnic groups of western Ethiopia.</i></p>
group of people	<p>social group  <i>People sharing some social relation</i></p> <p>people  <i>(plural) any group of human beings (men or women or children) collectively</i></p> <p>group  <i>Any number of entities (members) considered as a unit</i></p>
window	<p>window  <i>A framework of wood or metal that contains a glass windowpane and is built into a wall or roof to admit light or air</i></p>

Таблица 3

В Таблице 3 для значения выражения “group of people”

предложено общее значение, при этом, предложенное значение не полностью удовлетворяет контекст. «Группа людей», как было в примере 10, не обязательно должна включать в себя социальные связи между членами группы, в любом случае, контекст предложения это не предусматривает.

Невозможно оценить реализацию каждого из языков, доступных для обработки в системе, но русский язык работает не всегда должным образом. Встречаются опечатки, такие как “брег” вместо “берег”, неточности в выборе значений. Лучше реализованы такие языки, как немецкий (пример 11):

11) Gestern übernahm die päpstliche Armee Franz Kafkas Schloss.  
(пер. Вчера папская армия захватила замок Франца Кафки).

Найден ные сущности	Значения
Gestern	Vчера is a 1988 Bulgarian film.
übernah m	adopt <i>Take on titles, offices, duties, responsibilities</i>
päpstlic he	apostolic <i>Proceeding from or ordered by or subject to a pope or the papacy regarded as the successor of the Apostles</i>
Armee	army <i>The military forces of a nation</i>
Franz Kafkas	Franz Kafka <i>Czech novelist who wrote in German about a nightmarish world of isolated and troubled individuals (1883-1924)</i> Franz <i>Given name</i>
Schloss	castle <i>A large and stately mansion</i>

Таблица 4

В этом примере с существительным “Schloss” правильно сопоставляется английское слово “castle”, хотя у слова также есть омоним со значением “lock”. Для слова “gestern” вместо значения

«вчера» устанавливается название болгарского фильма, что нельзя считать удачной дизамбигуацией.

Возможна дизамбигуация как одного, так и нескольких предложений. В таком случае программа может учитывать расширенный контекст, когда при определении значения слова учитывается окружение не только в пределах предложения. Количество анализируемого материала неограниченно, дизамбигуируется сразу весь текст и его количество никак не влияет на скорость обработки. В похожем примере (14) со словом “nails”, как и в случае с алгоритмом Леска, было определено правильное значение “гвозди”, так как в следующем предложении в сочетании с указательным местоимением “them” использован глагол “hammer” - “забивать”.

14) I like covering the nails with bright red. And then I hammer them into the coffin.

В примере 14 слову “nails” установлено значение “horny plate covering and protecting part of the dorsal surface of the digits”, но, если увеличить окружение, добавив еще несколько предложений, значение может измениться. Прибавим к предыдущему примеру еще 2 предложения и получим текст: “I like covering the nails bright red. And then I hammer them into the coffin. I work in a beauty salon. I am a responsible employee, get along well with clients, do a very beautiful manicure.” Система снова правильно разрешит многозначность в слове “nail” и возвратит ему значение “ноготь”. Программа принимает такое решения из-за присутствия в добавленных предложениях таких словосочетаний, как “a beauty salon” и таких слов, как “manicure”.

Система не обновляется и не прорабатывается уже на протяжении нескольких лет. Такой вывод можно сделать из того, что она работает на базе семантической сети Babelnet 3.0, которая была выпущена в 2014 году. За последние годы эта семантическая сеть впоследствии получала множество модификаций, была обновлена до версии 4.0, а в данный момент получила live-версию, в которую постоянно загружаются новые синсеты. Возможно, этот факт является объяснением не самого лучшего уровня работы программы.

Для сравнения эффективности работы систем, их успешности в выполнении своих функций, было проведено статистическое

сравнение на небольшой выборке в 50 предложений. Использовались как предложения с частотной лексикой, так и более редкие в употреблении, наряду с предложениями с придуманным нами контекстом. В выборке, направленной на определение значения слова из контекста, присутствовали многозначные слова и омонимы, некоторые фразовые глаголы и идиомы. Предложения-примеры из предыдущих пунктов, с помощью которых демонстрировались возможности систем, тоже учитывались в тестировании. Предложения вводились в систему поочередно, и системами маркировалось ключевое слово (Алгоритм Леска) или все заданные слова (Babelfy) определенным значением или значениями. После того, как испытание было проведено, на основе полученных результатов вычислялся показатель эффективности работы систем. Число предложений  $N$  с верно определенными значениями делилось на общее число предложений  $N_0$ .

Тестирование проводилось для предложений на английском языке, что связано с низким качеством работы Babelfy с русским языком. Алгоритм Леска более применим на материале русского языка в связи с наличием тезаурусов, но доступной готовой реализации в виде модуля для него нет.

#### **Результаты и обсуждение**

Алгоритм Леска дал верный результат в 40 предложениях из 50. Следовательно, его коэффициент эффективности на основе данной выборки равен  $40/50=0,8$ . Система Babelfy сработала правильно в 36 случаях из 50, и ее коэффициент эффективности оказался равен 0,72.

Система	$N$	$N_0$	Коэффициент эффективности
LESK	4 0	50	0,8
Babelfy	3 6	50	0,72

*Таблица 5*

Алгоритм Леска оказался эффективнее несмотря на то, что был разработан значительно раньше системы Babelfy. В целом, данные показатели отражают эффективность работы данных систем в

определенной степени, и можно сказать, что они работают согласно своей цели - определение значения слова из его контекста. Но число ошибок и недочетов свидетельствуют об их несовершенстве и невысокой надежности. Тем не менее, нужно отметить факт сложной формализации семантики естественных языков, поэтому, вполне логично ожидать не самые высокие результаты. Даже носитель языка не всегда может верно определить значение многозначного слова в контексте, для автоматических систем это до сих пор открытая задача. Анализ проводился на выборке английских предложений, для других языков показатель эффективности был бы ниже, что связано как с большим количеством ресурсов для обработки английского языка, так и с более качественной предварительной обработкой английских слов, так как перед определением значения слово должно приводиться в нормальную форму.

Что касается удобства применения систем, сложно сделать однозначный вывод. Система Babelfy имеет демоверсию, более удобную для пользователей, не имеющих навыков программирования. Тем не менее, встроить ее в собственную систему обработки естественного языка сложнее. Алгоритм Леска имеет реализацию на базе библиотеки NLTK и тезауруса Wordnet, что позволит разработчикам исследовательских прототипов применять его как алгоритм для дизамбигуации в своих системах.

### **Выводы**

Дизамбигуация является важным компонентом лингвистических систем, который применяется в различных областях обработки естественного языка. Создание качественных и быстрых алгоритмов снятия многозначности до сих пор является востребованной задачей. Алгоритм Леска считается классическим подходом для дизамбигуации, так как он учитывает контекст слова, что, с позиции дистрибутивной семантики, является основополагающим для выделения значения. В настоящее время большую популярность получили алгоритмы и системы, работающие с данными Википедии, примером такой системы является Babelfy.

Было проведено статистическое сравнение работы систем на собранной выборке примеров, и большую эффективность показал алгоритм Леска. Обе протестированные системы способны

определять значения слов верно, но из-за наличия постоянных ошибок можно сделать вывод, что происходит это не всегда успешно.

Резюмируя проведенные исследования, стоит сказать, что системы дизамбигуации в состоянии определять значения слов из их контекста во множестве случаев, и, тем самым, способствовать пониманию языка и упрощать с ним работу. Однако системы не в состоянии показывать эффективность постоянно, и основной причиной этого является ограниченность библиотек и баз данных, на которых они основаны. Несмотря на ошибки, и алгоритм Леска, и система Babelfy способны снимать многозначность на уровне, достаточном для систем обработки естественного языка.

#### Литература

1. Алексеевский Д.А. Методы автоматического выделения тезаурусных отношений на основе словарных толкований: дисс. ... канд. филол.н. Москва, 2018.
2. Розенталь Д.Э., Теленкова М.А. Словарь-справочник лингвистических терминов. Пособие для учителя. – 3-е изд., испр. и доп. – М.: Просвещение, 1985.
3. Турдаков Д.Ю. Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов: дис. ... к. ф.-м. н. – М., 2010.
4. Ayetiran E.F., Agbele K. An optimized Lesk-based algorithm for word sense disambiguation // Open Computer Science. – 2018. – Vol. 8. – №. 1. – Pp. 165-172.
5. Booth A.D., Locke, W.N. (Eds.) Machine translation of languages: fourteen essays. – Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955. – Pp. 15-23.
6. Butnaru A.M., Ionescu R. T. Shotgun. WSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation // IEEE Access. – 2019. – Vol. 7. – Pp. 120961-120975.
7. Camacho-Collados J., Navigli R., Raganato A. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. – EACL. – 2017.
8. Iacobacci I., Navigli. R., Pilehvar M.T. Embeddings for Word Sense Disambiguation: An Evaluation Study. – ACL. – 2016.

9. Lenat D. Computers versus Common Sense (GoogleTechTalks on youtube)
10. Lesk M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone // Proceedings of the 5th Annual International Conference on Systems Documentation. – 1986. – C. 24-26.
11. Navigli. R. An up-to-date state of the art of the field // Word Sense Disambiguation: A Survey, ACM Computing Surveys. – № 41(2). – 2009. – Pp. 1-69.
12. Schütze H. 1998. Automatic word sense discrimination. Computational Linguistics. – № 24(1). – Pp. 97-123.
13. Weaver W. 1949. Translation. Machine Translation of Languages: Fourteen Essays. – Cambridge, MA: MIT Press.

#### References

- Alekseevskiy, D.A. (2018). *Metody avtomaticheskogo vydeleniya tezaurusnykh otnosheniy na osnove slovarnykh tolkovaniy*. [Methods of automatic allocation of thesaurus relations on the basis of vocabulary interpretations]. (Candidate thesis, Moscow, Russia).
- Ayetiran, E. F., & Agbele, K. (2018). An optimized Lesk-based algorithm for word sense disambiguation. *Open Computer Science*, 8(1), 165-172.
- Booth, A.D. & Locke, W.N (1955). (Eds.) *Machine translation of languages: fourteen essays*, (pp. 15-23). Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology.
- Butnaru, A. M., & Ionescu, R. T. (2019). ShotgunWSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation. *IEEE Access*, 7, 120961-120975.
- Camacho-Collados, J., Navigli, R., & Raganato, A. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. *EACL*, 99–110.
- Iacobacci, I., Navigli, R., & Pilehvar, M.T. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. *ACL*, 897–907.
- Lenat, D. (2019). *Computers versus Common Sense (GoogleTechTalks on youtube)*. Electronic resource. Retrieved from <https://www.youtube.com/user/GoogleTechTalks>

- Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, (pp. 24-26).
- Navigli, R. (2009). An up-to-date state of the art of the field. In *Word Sense Disambiguation: A Survey*. ACM Computing Surveys, 41(2), 1-69.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1), 97-123.
- Rosenthal, D.E., Telenkova, M.A. (1985). *Slovar-spravochnik lingvisticheskikh terminov [Dictionary of linguistic terms]*. – Moscow: Enlightenment.
- Turdakov, D. Yu. (2010). *Metody i programmnyye sredstva razresheniya leksicheskoy mnogoznachnosti terminov na osnove setey dokumentov [Methods and software to resolve the lexical ambiguity of terms based on networks of documents]*. (Candidate thesis, Moscow, Russia).
- Weaver, W. (1949). Translation. In Locke, W.N. and Booth, A.D. (Eds.) *Machine Translation of Languages: Fourteen Essays*. Cambridge, MA: MIT Press.

УДК 811.112

<https://doi.org/10.25076/vpl.36.04>

О.И. Максименко, Е.П. Подлегаева

Московский государственный областной университет

## ГЕНДЕРНЫЕ ОСОБЕННОСТИ ДЕТСКОЙ АНИМАЦИИ (ПРОБЛЕМЫ ПЕРЕВОДА НА ПРИМЕРЕ KIKORIKI И GOGORIKI)

В статье рассматривается проблематика локализации и адаптации мультисемиотических видеоверbalных анимационных текстов с русского языка на английский с учетом гендерных особенностей персонажей. Языковым материалом для анализа послужили авторские транскрипты серий российского мульти сериала «Смешарики», переведенного на британский вариант английского языка под названием *Kikoriki*. Интерес к межкультурной адаптации поликодового видеовербального