Часть III. «Исследования»

УДК 81.322.4
ББК 81.184

С. А. Лычко,
А. Е. Харламенков

Подход к разрешению лексической многозначности слова при переводе на русский жестовый язык

Московский политехнический университет
г. Москва

***Аннотация**: В статье проанализированы подходы к решению проблемы по разрешению лексической многозначности (WSD) с учётом контекста слова, приведены практические рекомендации по совместному использованию метода Леска и метода, основанного на семантической близости слов.*

***Ключевые слова**: искусственный интеллект, компьютерная лингвистика, машинный перевод, разрешение лексической многозначности, обработка естественного языка, WSD.*

Введение

В настоящее время существует огромное количество систем автоматического машинного перевода, предоставляющих довольно высокое качество перевода. Для качественного автоматического перевода недостаточно прямого перевода слов — необходимо строить синтаксически верные конструкции, учитывать идиомы и особенности языка и т.д. Одной из важнейших задач машинного перевода является задача разрешения лексической многозначности (WSD), которая заключается в выборе релевантного контексту значения многозначного слова или словосочетания.

Аналогичная проблема существует и в задачах перевода с жестового языка и наоборот. Попытке решения данной задачи для русского жестового языка на примере«Электронной справочно-аналитической системы “Толковый лексикографический словарь русского жестового языка”»[1] [1; 2; 6–9] и посвящается настоящая статья.

Поиск лучшего метода для разрешения лексической неоднозначности

Цель исследования. Анализ существующих подходов к разрешению лексической многозначности (WSD) с созданием комбинированного метода.

[1]Ранее данная система называлась «Электронная справочно-аналитическая система “Русско-жестовый толковый словарь”».

Задача исследования:

1. Проанализировать существующие подходы к разрешению лексической неоднозначности.
2. Выявить положительные и отрицательные стороны данных подходов.
3. Разработать рекомендации по совместному использованию выбранных методов.

Пример разрешения лексической неоднозначности в системе Google Translate:

"Этот *замок* сложно открыть" → "This *lock* is difficult to open

"*Замок* возвышался над озером" → "The *castle* towered over the lake".

Слово-омоним "замок" переводится разными словами на английский язык. Общая схема перевода слов-омонимов с учётом контекста представлена на рисунке 1.

Рисунок 1 - Схема разрешения лексической неоднозначности с использованием внешних знаний

Задачу разрешения лексической неоднозначности с учётом контекста можно представить следующим образом:

- На вход системы подаётся слово и его контекст. Под контекстом здесь понимается окно из нескольких слов вокруг целевого.
- Система классифицирует данное слово, используя его контекст, причём каждому возможному классу соответствует релевантное слово на языке, на который осуществляется перевод. В методах, основанных на внешних знаниях для этого обычно применяются контексты классов, представляющие собой словарные определения, семантически близкие слова, слова, часто встречающиеся в одном тексте со словом в значении, соответствующем данному классу и т. д.

Первые подходы к разрешению лексической неоднозначности начали появляться ещё в середине XX века; в настоящее же время количество методов решения этой задачи весьма велико. Ниже приводятся основные современные подходы к разрешению лексической неоднозначности [3, с. 30-46]:

1. Методы, основанные на внешних знаниях:
   - Метод Леска. Метрикой принадлежности слова к классу (значению) в данном методе является степень пересечения множества слов контекста и множества слов в словарных определениях, соответствующих данному значению слова [4].
   - Методы, основанные на семантической близости слов. Данная группа методов оперирует понятием семантической близости слов, под которой обычно понимается расстояние между двумя словами в определённых структурах данных, представляющих собою граф, где слова в узлах графа семантически связаны. Примером таких структур являются тезаурусы, подобные WordNet [5], или граф переходов проекта «Википедия» [3, с. 59-65].
   - К этой группе могут относиться разнообразные эвристики и методы, основанные на синтаксических зависимостях слова.
2. Методы, основанные на машинном обучении:
   - Обучение по размеченным корпусам. Данная группа подходов основана на применении разнообразных методов машинного обучения к размеченных вручную корпусам.
   - Обучение по неразмеченным корпусам. Данная группа методов не использует внешних данных; определённые методы этой группы могут кластеризовать значения слов по контекстам и определить принадлежность слова к одному из кластеров. Недостатком методов данной группы является невысокая точность.

Сравнение подходов к разрешению лексической неоднозначности приведено в таблице 2.

Таблица 2 — Сравнение подходов к разрешению лексической неоднозначности.

	Метод Леска	**Методы, основанные на семантической близости**	**Обучение по размеченным корпусам**	**Обучение по неразмеченным корпусам**
Точность	Невысокая	Высокая	Очень высокая	Низкая
Требования к внешним данным	Средние	Высокие	Высокие	-
Тип внешних данных	Словарные определения	Тезаурусы, сети документов и т.д.	Размеченные вручную корпуса текстов	Любые тексты

На практике можно столкнуться с тем, что доступные внешние данные ограничены или содержат узкоспециализированную группу слов. Для увеличения точности распознавания и расширения доступной лексики можно использовать комбинированные методы. Возможный способ комбинации разных подходов представлен ниже:

1) Обработка контекста слова:

a. Выбор окна контекста вокруг целевого слова.

b. Удаление из контекста предлогов, союзов, местоимений — данные части речи не несут существенной семантической нагрузки.

c. Леммизация слов — приведение в начальную форму.

2) Следует понимать, что использование такой обработки исключает возможность применения методов, оперирующих синтаксическими связями в контексте. Данные методы необходимо использовать до такой обработки.

3) Применение метода Леска для нахождения оценок принадлежности слова к одному из классов.

4) Применения метода, использующего семантическую близость (например, сети документов [3, с. 49]) для нахождения оценок принадлежности слова к одному из классов.

5) Нормализация этих оценок, то есть приведение к диапазону [0, 1].

6) Итоговые оценки будут вычисляться как взвешенная сумма оценок, использующих разные методы.

7) Присвоение слову значения класса с наилучшей оценкой.

Заключение

Проанализировав подходы к разрешению лексической многозначности (WSD), описаны преимущества и недостатки существующих подходов. На основании полученных выводов, в качестве решения обозначенной проблемы для русского жестового языка на примере «Электронной справочно-аналитической системы "Толковый лексикографический словарь русского жестового языка"»[1] [1; 2; 6—9], предлагается использовать комбинацию метода Леска и дистрибутивно-семантического метода в условиях недостатка внешних данных или их недостаточного качества.

Литература

1. Харламенков А.Е. Аналитический обзор электронных on-line словарей жестовых языков: монография / А.Е. Харламенков. — Москва: РУСАЙНС, 2017. — 218 с.
2. Харламенков А.Е. Русский жестовый язык. Начала : [Учебное пособие] : Русский жестовый язык. Начала / А.Е. Харламенков. — Москва: Издательство «ОнтоПринт», . — 164 с.
3. Турдаков Д.Ю. Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов : Диссертация ... кандидата физико-математических наук ; 05.13.11 / Д.Ю. Турдаков. — Москва: Московский государственный университет имени М. В. Ломоносова, 2010. — 138 с.
4. Lesk M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone / M. Lesk // Proceedings of the 5th Annual International Conference on Systems Documentation : SIGDOC '86 / event-place: Toronto, Ontario, Canada. — New York, NY, USA: ACM, 1986. — Automatic Sense Disambiguation Using Machine Readable Dictionaries. — С. 24—26.

[1]Ранее данная система называлась «Электронная справочно-аналитическая система "Русско-жестовый толковый словарь"».

5. Princeton University. WordNet | A Lexical Database for English [Электронный ресурс]. — URL: https://wordnet.princeton.edu/ (дата обращения: 28.03.2019).
6. Харламенков А.Е. Методика преодоления безусловного рефлекса при постановке рук в процессе освоения дактильной и жестовой речи / А.Е. Харламенков // Научные труды Института непрерывного профессионального образования. — 2014. — Вып. № 3. — С. 44-49.
7. Харламенков А.Е. Резолюция Научно-практической конференции «Перспективы выхода из сложной ситуации с русским жестовым языком в сфере российского образования ввиду принятия ГОСТ Р 57636— 2017» / А.Е. Харламенков // Научные труды ЦНИИ русского жестового языка // Материалы конференции Научно-практическая конференция «Перспективы выхода из сложной ситуации с русским жестовым языком в сфере российского образования в виду принятия ГОСТ Р 57636— 2017» / ред. В.В. Кузьмин. — Москва: ЦНИИ русского жестового языка, 2018. — Т. № 1. — С. 19-28.
8. Харламенков А.Е. Создание «Электронной справочно-аналитической системы “Русско-жестовый толковый словарь”» Монография / А.Е. Харламенков // Научные труды Института Непрерывного Профессионального Образования. No 7. Монографические исследования / ред. Под научн. редакцией проф. П. С. Гуревича и проф. С. В. Чернова. С. 97-186. — Москва: Издательство Института Непрерывного Профессионального Образования, 2016. — С. 89.
9. Харламенков А.Е. Электронная справочно-аналитическая система «Русско-жестовый толковый словарь» / А.Е. Харламенков // Научные труды Института непрерывного профессионального образования : Материалы Общероссийской научно-практической конференции «Наука. Образование. Проектная деятельность: Россия — XXI век» / ред. С.В. Чернов. — Москва: Институт Непрерывного Профессионального Образования, 2014. — Вып. № 3. — С. 24-43.