ST. PETERSBURG STATE UNIVERSITY
INSTITUTE FOR LINGUISTIC STUDIES (RAS)
HERZEN STATE PEDAGOGICAL UNIVERSITY OF RUSSIA

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
«CORPUS LINGUISTICS – 2015»

June 22–26, 2015, St. Petersburg

St. Petersburg
2015

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ РАН
РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ
им. А.И. ГЕРЦЕНА

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
«КОРПУСНАЯ ЛИНГВИСТИКА – 2015»

22–26 июня 2015 г., Санкт-Петербург

Санкт-Петербург
2015

Э.С. Клышинский, В.К. Логачева, Й. Нивре
E.S. Klyshinsky, V.K. Logacheva, J. Nivre

РАСПРЕДЕЛЕНИЕ НЕОДНОЗНАЧНЫХ СЛОВ В НЕКОТОРЫХ ЕВРОПЕЙСКИХ ЯЗЫКАХ[1]

THE DISTRIBUTION OF AMBIGUOUS WORDS IN EUROPEAN LANGUAGES

Аннотация. Результаты морфологического анализа часто неоднозначны - одно словоупотребление может быть формой нескольких разных слов с разными грамматическими параметрами. В статье рассматривается классификация типов такой неоднозначности и вероятностное распределение словоупотреблений по типам для ряда языков. Эксперименты показывают, что форма распределения различается от языка к языку, но сохраняется для корпусов разных типов, а также устойчива к изменению параметров грамматического словаря.

Ключевые слова. Грамматическая неоднозначность, языковая сложность, вероятностное распределение.

Abstract. During the tagging process, a word can be tagged with several parts of speech and/or sets of grammatical parameters. We classify the types of this ambiguity and show the probability distributions of words among the ambiguity types for a number of languages. According to our experiments, the shape of this distribution does not depend on type of a corpus and variations in the set of used grammatical parameters.

Keywords. Grammatical ambiguity, language complexity, statistical distribution.

[1] Данная работа выполнена при финансовой поддержке гранта РГНФ 15-04-12019.

1. Введение

Одним из вопросов в автоматической обработке текстов является перенос методов и техник на новые языки. Несмотря на то, что многие из методов декларируются как языконезависимые, они редко тестируются больше, чем на нескольких языках, а анализ применимости данных методов к другим языкам с учетом их особенностей не проводится. Чаще всего подобные ограничения выявляются для методов и систем, изначально ориентированных на применение для английского языка, но применяемых к языкам с высокой степенью флексии, например, к русскому языку. Так, например, Протопопова и Бочаров в работе [Protopopova, Bocharov 2013] описывают применение метода снятия омонимии Брилля к русскому языку. Хотя сам по себе метод заявлен как применимый к любому языку, авторам пришлось учитывать флексию русского языка, так как правила английского языка выдавали низкий процент точности.

Задачей данной работы была количественная оценка различий в статистике распределения грамматически неоднозначных слов в тексте по их грамматическим характеристикам (части речи, лемме и грамматическим параметрам). Эксперименты были проведены для английского, французского, испанского, итальянского, немецкого, русского и польского языков.

2. Существующие решения

В большей части работ, связанных с разрешением грамматической неоднозначности, приводится информация лишь для одного языка. В общем случае, темой таких исследований является частное исследование, например, распределение падежей существительных в текстах некоторого жанра [Lyashevskaya 2013] и др. Анализ литературы позволяет найти рабо-

ты о частеречной омонимии в чешском, английском, венгерском, эстонском, литовском и латышском языках.

Рассматриваемые в здесь вопросы в значительной мере коррелируют с вопросами языковой сложности. Однако не все работы в этой области посвящены количественным оценкам. Так, World Atlas of Language Structures Online (WALS, http://wals.info) содержит в основном качественную информацию. При этом некоторые существующие математические методы оценки языковой сложности вызывают определенные сомнения. Так, в работе [Sadeniemi 2008] (как и в некоторых других) проводится оценка колмогоровской сложности текста оценкой степени его архивации, при том, что не совсем ясно, сложность чего именно замеряется таким образом. При этом там же сложность считается как функции от числа выделенных в тексте морфем.

Краткое сравнение грамматической неоднозначности слов для русских и английских текстов мы проводили в [Клышинский 2013], однако двух языков недостаточно для полноценного сравнения. В данной работе мы включили в рассмотрение еще несколько языков из разных языковых групп.

3. Метод анализа омонимичной лексики и результаты экспериментов

Задачей морфологического анализа слова является определение его леммы (начальной формы), его части речи и набора грамматических параметров. Если слово отсутствует в словаре, назовем его несловарным - для такого слова анализ не может быть проведен. Если же слову в результате анализа было приписано более одной начальной формы, леммы или набора параметров, назовем его неоднозначным. Таким образом, по результатам анализа мы разделили слова на шесть типов: однозначное, неоднозначное по параметрам, неоднозначное по части речи, неоднозначное по лемме, неоднозначное по части речи и лемме, несловарное.

Мы рассчитали распределение словоупотреблений по этим типам для семи европейских языков: английский, русский, французский, испанский, итальянский, немецкий и польский. Для того чтобы устранить влияние стиля корпуса, мы проводили эксперименты только на новостных корпусах. Характеристики корпусов и теггеров, использованных для разметки, приведены в табл. 1. Статистика распределения словоупотреблений по типам омонимии для анализируемых языков представлена на рис. 1 (более подробно результаты изложены в препринте http://library.keldysh.ru/preprint.asp?id=2015-4).

Таблица 1. Характеристики систем разметки и корпусов

Язык	**Система разметки**	**Корпус**	**Размер корпуса**
Англ.	Расширенная AOT.ru	Reuters	300 млн
Франц.	Morphalu	Le Parisien	43.1 млн
Исп.	FreeLing	Abc.es	15.2 млн
Итальян.	FreeLing	Corriere dela Serra	7.9 млн
Нем.	TreeTagger	Die Zeit	7.1 млн
Русск.	Расширенная AOT.ru	Lenta.ru	32.4 млн
Польск.	Morfologik	Different sources	21.2 млн

Сравнение распределения слов по типам омонимии дает нам информацию о том, насколько похожи языки с точки зрения количества и типа морфологической неоднозначности. Распределения для немецкого, итальянского и испанского языков близки между собой: корреляция составила от 0,83 до 0,91, а для французского и итальянского – 0,93. Распределения слов по типам омонимии в романских и немецком языках похожи, английский и русский отличаются от них, а польский не коррелирует ни с одним языком, кроме слабого сходства с русским.

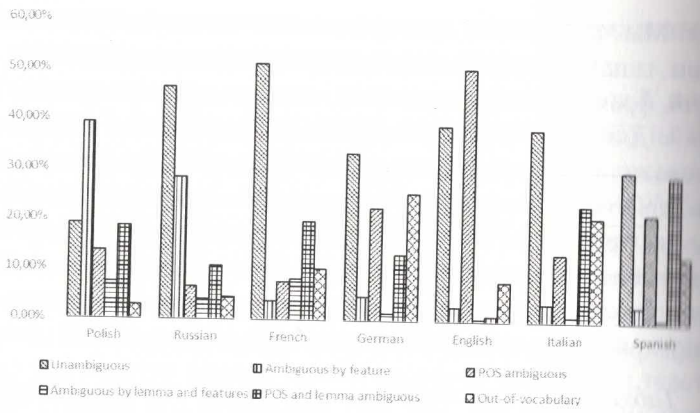

Рис. 1. Распределение словоупотреблений по типам неоднозначности для различных языков

В некоторых задачах можно использовать лишь слова, однозначные по части речи. В русском языке их более 80%, в польском и французском менее 70%, в английском и испанском около 40% и 45%. Значит, при анализе английских и испанских текстов проблему составляют слова, омонимичные по части речи; для русского и польского языка проблему составляет неоднозначность параметров. То есть, богатая флексия языка требует для машинного обучения размеченных корпусов большего размера, но позволяет в ряде задач (например, при применении подхода «мешок слов») обойтись без снятия омонимии.

В экспериментах мы использовали разнородные ресурсы и инструменты. Чтобы показать отсутствие влияния такой неоднородности на распределение неоднозначных слов по типам мы провели дополнительные эксперименты. Слова в словарях сортировались по частоте употребления в корпусах и расчет проводился для 1000 самых частотных слов. Эксперименты показали, что изменение размеров словаря не изменяет формы распределения. Корреляция распределений русского языка была не ниже 0,993, для английского – 0,999. Таким образом,

форма распределения определяется наиболее частотными словами языка.

Мы провели анализ параллельных (французский, немецкий и испанский) корпусов (http://www.statmt.org/wmt13/translation-task.html#download). Для французского и испанского языков распределения получились сходными (корреляция для французского корпуса – 0,999, испанского – 0,998, немецкого – 0,94). Также результаты экспериментов были повторены на корпусе Син-ТагРус. Полученное распределение имеет сходную форму, суммарная разница в распределениях находится на уровне 10%, следовательно, распределение зависит от стиля текста или особенностей выбранного словаря, однако его форма сохраняется.

Литература

1. *Клышинский Э.С., Кочеткова Н.А, Мансурова О.Ю., Ягунова Е.В., Максимов В.Ю., Карпик О.В.* (2013), Формирование модели сочетаемости слов русского языка и исследование ее свойств // Препринты ИПМ им. М.В. Келдыша. 2013. № 41. 23 с. URL: http://library.keldysh.ru/preprint.asp?id=2013-41

2. *Lyashevskaya O.* (2013), Frequency Dictionary of Inflectional Paradigms: Core Russian Vocabulary. // Preprints of HSE, Series: Humanity, WP BRP 35/HUM/2013, available at http://www.hse.ru/data/2013/06/27/1285976210/35HUM2013.pdf

3. *Protopopova E.V., Bocharov V.V.* (2013), Unsupervised learning of part-of-speech disambiguation rules // In Proc. of Computational Linguistics and Intellectual Technologies (Dialog 2013). Bekasovo, Russia.

4. *Sadeniemi M., Kettunen K., Lindh-Knuutila T., Honkela T.* (2008), Complexity of European Union Languages: A Comparative Approach // Journal of Quantitative Linguistics. Vol.15, N2, pp. 185-211.

References

1. *Klyshinsky E.S., Kochetkova N.A., Mansurova J/Yu., Iagunova E.V., Maksimov V. Yu., Karpik O.V.* (2013), Formirovanie modeli sochetaemosti slov russkogo jazika i issledovanie eje svojstv [Analysis of Words' Ambiguity in European Languages] // Preprinti IPM imeni M.V. Keldisha [Preprints of Keldysh IAM] № 41. 23 p. URL: http://library.keldysh.ru/preprint.asp?id=2013-41

2. *Lyashevskaya O.* (2013), Frequency Dictionary of Inflectional Paradigms: Core Russian Vocabulary. // Preprints of HSE, Series: Humanity, WP BRP 35/HUM/2013, available at http://www.hse.ru/data/2013/06/27/1285976210/35HUM2013.pdf

3. *Protopopova E.V., Bocharov V.V.* (2013), Unsupervised learning of part-of-speech disambiguation rules // In Proc. of Computational Linguistics and Intellectual Technologies (Dialog 2013). Bekasovo, Russia.

4. *Sadeniemi M., Kettunen K., Lindh-Knuutila T., Honkela T.* (2008), Complexity of European Union Languages: A Comparative Approach // Journal of Quantitative Linguistics. Vol. 15, N 2, pp. 185-211.

Клышинский Эдуард Станиславович
Национальный исследовательский университет «Высшая школа экономики», МИЭМ (Россия).
Klyshinsky Edward
National Research University "Higher School of Economics», MIEM (Russia).
E-mail: eklyshinsky@hse.ru
Логачева Варвара Константиновна
Университет Шеффилда (Великобритания).
Logacheva Varvara
University of Sheffield (Great Britain).
E-mail: v.logacheva@sheffield.ac.uk
Нивре Йуаким

Университет г. Уппсала (Швеция).
Nivre Joakim
Uppsala University (Sweden).
E-mail: joakim.nivre@lingfil.uu.se

Научное издание

Труды международной конференции «Корпусная лингвистика – 2015»

Сборник издается в авторской редакции

Компьютерная подготовка издания
А.Сумбатовой

Подписано в печать с оригинала-макета заказчика 11.06.2015.
Формат 60x84/16. Печать офсетная. Усл. печ. л. 27,0.
Тираж 120 экз. Заказ № 3841

Издательство СПбГУ. 199004, Санкт-Петербург, В.О., 6-я линия, 11/21.
Тел. (812)328-96-17; факс (812)328-44-22
E-mail: editor@unipress.ru
www.unipress.ru

Типография ООО «Литография».
191119, С.-Петербург, Днепропетровская ул., 8.