5. SHO‘BA. TABIIY TILGA AVTOMATIK ISHLOV BERISH MUAMMOLARI

ПРИМЕНЕНИЕ МЕТОДОВ NLP В КОРПУСНЫХ ИССЛЕДОВАНИЯХ: ОСОБЕННОСТИ И ОГРАНИЧЕНИЯ

НИГМАТОВА Лолахон
DSc, доцент кафедры русского литературоведения
Бухарского государственного университета
nigmatovalolaxon@gmail.com
АВЕЗОВ Сухроб
преподаватель кафедры русского литературоведения
Бухарского государственного университета
1990senigama@gmail.com

Аннотация. Обработка естественного языка (NLP) – анализ человеческого языка компьютерными программами. NLP включает задачи от простых (разделение текста на слова) до сложных (преобразование речи в текст, аннотация синтаксическими характеристиками). NLP способствует развитию корпусной лингвистики и машинному обучению. Однако, NLP имеет слабые стороны и потенциальные ограничения. Важно осознавать возможности и недостатки инструментов NLP для корректного использования. В статье рассматриваются NLP-процессы (токенизация, лемматизация, маркировка частей речи, анализ составляющих и анализ зависимостей) для повышения грамотности в области NLP.

Ключевые слова: NLP, корпусное исследование, токенизация, лемматизация, аннотация частей речи, аннотация составляющих, аннотация зависимостей, обучающие данные, анализ языковых данных.

Abstract. Natural Language Processing (NLP) is the analysis of human language by computer programs. NLP includes tasks ranging from simple (e.g., splitting text into words) to complex (e.g., transforming speech into text, annotating syntactic features). NLP contributes to the development of corpus linguistics and machine learning. However, NLP has weaknesses and potential limitations. It is essential to recognize the capabilities and drawbacks of NLP tools for their proper use. The article discusses NLP processes (tokenization, lemmatization, part-of-speech tagging, constituency analysis, and dependency analysis) to increase literacy in the field of NLP.

Keywords: NLP, corpus research, tokenization, lemmatization, part-of-speech annotation, component annotation, dependency annotation, training data, language data analysis.

Область NLP довольно обширна и включает в себя множество процессов. В этой статье основное внимание уделяется пяти NLP-анализам, которые относительно часто используются в исследованиях корпусов, а

именно: токенизации, лемматизации, аннотации частей речи, аннотации составляющих и аннотации зависимостей. Хотя другие NLP-анализы, такие как семантика векторного пространства, также имеют возможности для исследований корпусов, в нашей работе акцент делается на лингвистической аннотации.

Анализы NLP опираются на закономерности в языковых данных, на которых они обучаются [Polio, 2018: 179]. Лингвистические особенности, которые явно кодируются и используются с небольшой двусмысленностью в языковом корпусе, будут автоматически аннотированы с гораздо большей точностью, чем те, которые менее явно кодируются и/или используются двусмысленно. Кроме того, степень, в которой использование определенной языковой особенности в обучающих данных представляет собой область использования целевого языка, повлияет на точность автоматической аннотации. По крайней мере, две характеристики обучающих данных будут влиять на репрезентативность, а именно размер корпуса и степень сравнимости регистра обучающего корпуса и целевого корпуса.

Таким образом, анализы NLP зависят от характеристик обучающих данных и репрезентативности этих данных относительно целевого корпуса. Размер и сходство регистров обучающего и целевого корпусов влияют на точность автоматической аннотации, и эти факторы могут быть существенными при рассмотрении анализа языковых данных учащихся. Важно учитывать эти различия и особенности при использовании NLP-инструментов для анализа корпусов и других языковых данных.

Токенизация заключается в разделении текста на словесные единицы. В узбекском и других языках, где словесные единицы разделяются пробелами, токенизация является относительно простой и высокоточной задачей, включающей два основных этапа. Во-первых, большинство (если не все) знаков препинания должны быть отделены от словесных единиц. Поскольку некоторые знаки препинания могут использоваться неоднозначно, токенизаторы могут использовать статистические/машинные модели обучения для точного разделения несловесных знаков препинания от слов. Во-вторых, слова необходимо разделить с использованием пробелов, а все несловесные элементы могут быть удалены в зависимости от целей исследователя. Для большинства текстов токенизацию можно выполнить с высокой степенью точности, хотя опечатки могут вызвать ошибки.

В языках, где словесные единицы не обязательно разделяются пробелами и имеют неоднозначные флективные морфемы (например, корейский), токенизация слов может быть менее простой и стать источником ошибок в анализе корпусов [Авезов, 2023: 54].

Лемматизация включает группировку слов в их флективных формах (например, «бежал») по их неизмененным формам (например, «бежать»), чтобы их можно было анализировать как единую словоформу. Хотя лемматизация не требуется и не обязательно предпочтительна во всех ситуациях или для всех языков, она часто используется во многих

исследованиях языкового корпуса. Распространенный метод лемматизации текста заключается в использовании списка лемм, основанных на поверхностной форме [Khamidovna, 2023: 240].

Для необработанных текстов, опечатки и ошибки в написании слов снижают точность лемматизации, что, в свою очередь, может повлиять на последующие лингвистические анализы.

Аннотация частей речи предоставляет множество возможностей для исследователей корпусов. Как было отмечено в предыдущем разделе, аннотация частей речи может использоваться для разрешения омонимии, но она также может использоваться для проведения лексико-грамматических анализов языкового использования. Кроме того, аннотации частей речи служат основой для сложной синтаксической аннотации, такой как аннотация составляющих и зависимостей.

Существует несколько конкретных подходов к автоматической разметке частей речи, которые отличаются по используемым наборам признаков и статистическим/машинным алгоритмам обучения. Однако большинство разметчиков аннотации частей речи применяют одинаковый базовый подход. Во-первых, всем словам с однозначными метками частей речи в обучающих данных присваиваются соответствующие метки. Затем ряд контекстуальных особенностей, таких как метка аннотации частей речи предыдущего слова или слов, окончания целевого слова и предыдущего слова или слов, само целевое слово и т. д., используются в качестве предикторов в статистическом или машинном алгоритме обучения для предсказания частей речи слов с неоднозначными метками в обучающих данных или отсутствующих в обучающих данных. Разметчики могут достигать высокой точности аннотации как для хорошо отредактированных текстов на родном языке, соответствующих области(ям) использования языка обучающего корпуса, так и для многих типов текстов на втором языке.

Как и в случае с токенизацией и лемматизацией, опечатки и орфографические ошибки, а также специфические для языка проблемы могут вызвать ошибки в аннотации частей речи. Кроме того, поскольку разметка частей речи основана на закономерностях в последовательности слов, порядок слов и коллокационные ошибки в текстах учащихся могут повлиять на точность аннотации частей речи для слов, чья поверхностная форма может быть присвоена нескольким меткам.

Синтаксические парсеры – конституенты создают синтаксические деревья конституентов для предложений в тексте. Одно из популярных применений автоматической аннотации синтаксического разбора – расчет мер синтаксической сложности, таких как средняя длина T-юнита [Lu, 2015: 20].

Синтаксические парсеры – конституенты используют фразовые структурные правила, сгенерированные из обучающих корпусов, чтобы создать деревья синтаксических конституентов на уровне предложений. Сначала тексты размечаются частями речи, затем эти метки используются

совместно с фразовыми структурными правилами для создания конкурирующих деревьев разбора на уровне предложений. Наконец, статистические/машинные алгоритмы обучения используются для выбора наиболее вероятного дерева разбора для предложения.

В данной статье рассмотрены пять ключевых аспектов анализа NLP, которые находят широкое применение в исследованиях корпусов, а именно: токенизация, лемматизация, аннотация частей речи, аннотация составляющих и аннотация зависимостей. Качество и точность аннотации в NLP-анализах во многом определяются характеристиками обучающих данных и их репрезентативностью относительно целевого корпуса.

Важно учитывать факторы, такие как размер корпуса и степень сравнимости регистра обучающего корпуса и целевого корпуса, поскольку они влияют на точность автоматической аннотации. Опечатки, орфографические ошибки и специфические для языка проблемы могут вызывать ошибки в аннотации частей речи, лемматизации и токенизации.

Синтаксические парсеры, основанные на аннотации составляющих и зависимостей, также зависят от точности аннотации частей речи и адекватности фразовых структурных правил, полученных из обучающих корпусов. Поэтому важно тщательно выбирать обучающие данные и учитывать особенности целевого корпуса при использовании NLP-инструментов для анализа языковых данных.

В целом, NLP-анализы являются мощным инструментом для исследования корпусов и других языковых данных, но необходимо учитывать ограничения и потенциальные источники ошибок, связанные с характеристиками обучающих данных и репрезентативностью этих данных относительно целевого корпуса. Осознание этих факторов и применение адекватных методов обучения и аннотации способствует повышению точности анализа и дает исследователям возможность делать более обоснованные выводы на основе своих корпусных исследований.

Использованной литературы

1. Green C. Enriching the academic wordlist and Secondary Vocabulary Lists with lexicogrammar: Toward a pattern grammar of academic vocabulary //System. – 2019. – Т. 87. – С. 102–158.

2. Polio C., Yoon H. J. The reliability and validity of automated tools for examining variation in syntactic complexity across genres //International Journal of Applied Linguistics. – 2018. – Т. 28. – №. 1. – С. 165-188.

3. Авезов С., Юсупова А. Процесс обработки узбекского параллельного корпуса в условиях недостаточности данных //Евразийский журнал социальных наук, философии и культуры. – 2023. – Т. 3. – №. 3. – С. 49-58.

4. Khamidovna N. L. Expression of the Harmony of Language and Culture in World and Uzbek Lexicography //resmilitaris. – 2023. – Т. 13. – №. 1. – С. 233-244.

5. Lu X., Ai H. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds //Journal of second language writing. – 2015. – T. 29. – C. 16-27.