Анализ релевантных признаков для автоматического определения сложности русского текста как иностранного

Лапошина А.Н. (antonina.laposhina@gmail.com)
Кафедра компьютерной лингвистики Института лингвистики РГГУ, Москва

This article presents the results of a work on selecting and analyzing the linguistic features that may affect the Russian second language reading difficulty. First, basic approaches to readability measurement are discussed. Then we describe shortly the main differences from the first language text complexity problem. Next, we dwell in detail on the stage of collecting features and evaluation they correlation with reading difficulty. The results of this analysis can be used as a theoretical basis for studying second language texts complexity, and can also help tuning the machine learning model.

Key words: сложность текста, удобочитаемость, readability, text complexity, reading difficulty, computational linguistics

Введение

Чтение считается одним из важнейших аспектов методики обучения иностранному языку: раздел, посвященный чтению, входит во все сертификационные тесты, одной из необходимых составляющих понятия владения иностранным языком является способность читать и понимать неадаптированные тексты на этом языке. Но для достижения хорошего результата обучения текстовый материал должен подходить учащимся по уровню сложности грамматических конструкций, объему, лексическому наполнению и т.д. Подбор текстов для учебных пособий, поиск интересных неадаптированных текстов для занятий или самостоятельного чтения - всё это делает исследования в области автоматического определения сложности текстов актуальными и практически применимыми.

Формально сложность текста можно обозначить как сумму всех элементов текста, влияющих на понимание темы, скорость чтения и уровень интереса к прочитанному. В широком понимании этого термина на сложность текста могут влиять такие факторы как шрифт, сопровождение иллюстрациями, близость темы читателю, и эта тема также очень интересна, однако мы в своей работе будем оперировать более узким пониманием сложности текста, зависящей лишь от лингвистических факторов, самого текстового наполнения.

Определение сложности иностранного текста является частью более широкой проблемы автоматического определения сложности текста вообще. Мы предполагаем, что основными особенностями определения сложности текста как иностранного являются: более сильная зависимость от лексики и грамматики (так, например, Heilman et al. в своей работе [1] приводят результаты эксперимента, в ходе которого добавление грамматических признаков принесло бóльший прирост точности в коллекции текстов как иностранных - 22% против 7% как родных). Надо отметить, что данное исследование проводилось для английского языка и требует проверки для текстов на русском языке. Ещё одной важной особенностью проблемы измерения сложности текстов как иностранных является наличие единой понятной шкалы уровней владения языком как

иностранным и официальных документов, регламентирующих признаки текста того или иного уровня (государственные стандарты и лексические минимумы). Всё это будет нами учтено на этапе сбора признаков.

Подходы к изучению сложности текста

За более чем 70-летнюю историю изучения возможностей измерения сложности текста сложилось два основных подхода к автоматическому определению сложности текста:

1. Традиционный, основанный на различных формулах удобочитаемости.

Первые формулы на основе простых вычислений, таких как средняя длина слова, средняя длина предложения, количество длинных слов и т.д. появились еще в 30-гг. прошлого века. Среди наиболее популярных можно назвать формулы Флэша-Кинсайда, Дэйла-Чалл, SMOG (Simple Measure of Gobbledygook). Большинство таких формул были разработаны для английского языка; результатом их работы является число, обозначающее класс или возраст, на который рассчитан этот текст. Поэтому для определения сложности русского текста необходим пересчет параметров: здесь следует упомянуть работу И.В.Оборневой [9] и проект Readability.io Ивана Бегтина. Плюсом таких формул, безусловно, является простота и быстрота вычисления, однако они основаны лишь на поверхностных характеристика текста и не учитывают непосредственного наполнения текста.

2. Вычисление сложности на основе машинного обучения.

Проблема автоматического определения сложности текста может быть решена также в рамках машинного обучения. Это классическая задача построения предсказательной модели на основании обучения на тренировочном корпусе текстов и наборе признаков.

Данная работа выполнена в рамках второго подхода: следовательно, перед нами стояли задачи сбора корпуса текстов, поиска релевантных признаков на основе полученного корпуса и обучение модели машинного обучения на этих признаках. Далее мы кратко отразим первый и третий пункты и подробнее остановимся на этапе сбора и анализа признаков текста.

Выбор шкалы и сбор корпуса

В нашем исследовании предлагается принять за шкалу сложности текстов их соответствие общепринятым в методике уровням владения иностранным языком. Российская многоуровневая система тестирования включена в европейскую структуру языкового тестирования ALTE и содержит 6 уровней: A1, A2, B1 ,B2, C1, C2, от элементарного (A1) до уровня разрешения на преподавательскую деятельность на русском языке (C2).

Поскольку подобных готовых корпусов не представлялось возможным найти, было принято решение собрать собственный небольшой корпус. Он представляет собой около 600 текстов, взятых из учебной текстотеки ЦМО МГУ и учебных пособий, в методической справке которых был указан уровень владения языком, для которого он предназначается.

Для удобства пользования корпусом мы присвоили каждому уровню числовой эквивалент от 1 до 6 (A1 = 1, A2 = 2 и т.д.)

Сбор и поиск релевантных признаков

Дальнейшим и ключевым этапом нашего исследования является анализ полученного корпуса и сбор текстовых признаков. Для решения этой задачи был написан программный код на языке Python: текст проходит стадии деления на предложения, токенизации, морфологического анализа с помощью модуля Mystem и, непосредственно, подсчета признаков.

Можно выделить следующие классы интересующих нас признаков:

- ***Традиционные метрики текстов*** (такие как средние и медианные длины слов и предложений, процент слов длиннее 4 слогов и др.).

- ***Признаки на основе формул читабельности.*** Нами были выбраны вслед за И.Бегтиным пять наиболее широко использующихся в англоязычном мире формул для оценки сложности текстов: формула Флэша-Кинкайда, Колман-Лиау, Дэйла-Чалл, SMOG и Automated Readability Index. Поскольку результатом работы этих формул должно стать число, соответствующее возрасту и классу учащихся, для которого он предназначен, в формулах подбираются необходимые коэффициенты. Мы же перед собой такой задачи не ставим, наш интерес составляет именно новый признак, полученный отношением простейших метрик текста между собой, поэтому мы не используем коэффициенты.

- ***Лексические признаки.*** Тут в качестве признака выступает доля слов в тексте, относящихся к определенным интересующим нас спискам: первая подгруппа - это лексические минимумы, т.е. зафиксированный методистами список слов, которые студент должен знать на определенном уровне. Вторую подгруппу лексических признаков составляет доля слов, входящих в списки Частотного словаря современного русского языка (на материалах Национального корпуса русского языка) [8].

- ***Грамматические признаки.*** Для подсчета грамматических признаков была использована программа Mystem. Считалась доля того или иного грамматического признака 1) во всем тексте 2) в предложении.

- ***Семантические признаки.*** Ещё в работе Ю.А. Томиной [10], посвященной объективной оценке языковой трудности текстов, высказывалась мысль, что одним из факторов, влияющим на трудность текста, могут быть показатели абстрактности. Взяв за основу её предположение, мы использовали списки слов из семантической иерархии ABBYY COMPRENO с семантемами (своеобразными семантическими метками), характеризующими существительные с точки зрения абстрактности/конкретности. Получилось 4 списка:
  'lex_physical' содержит существительные, обозначающие конкретные материальные объекты, включая людей (e.g. 'котлета', 'стол', 'мама'); 'lex_virtual' - конкретные, но

виртуальные, нематериальные объекты (e.g. 'база', 'интернет'), 'lex_abstract' - самый большой список, со держащий в себе различные абстрактные понятия, включая термины (e.g. 'авангардизм', 'блажь', 'сглаживание') , и, наконец, в 'lex_substance' собраны обозначения субстанций (e.g. 'серебро', 'уксус').

Далее нам необходимо было измерить для каждого признака степень его корреляции со сложностью текста. В качестве меры был использован коэффициент корреляции Пирсона. Эта величина изменяется от -1 до +1. Чем она ближе к нулю, тем меньше наблюдается связь признака со сложностью. Положительный коэффициент говорит о положительной корреляции (с ростом признака растет и уровень сложности текста), отрицательный – наоборот (с ростом признака сложность текста падает). Расчеты проводились с помощью модуля для статистических расчетов scipy.stats на Python.

Признак	Коэффициент корреляции
mean_len_sentence	0.6541
median_len_sentence	0.6157
percent_of_long_words	0.5932
mean_len_word_in_syllables	0.5490
words	0.5369
mean_len_word	0.5332
median_punct_per_sentence	0.4731
sentences	0.4150
median_len_word	0.3707

Таблица 1. Степень корреляции традиционных метрик текста

Признак	Коэффициент корреляции
formula_smog	0.6863
formula_senter	0.6653
formula_flesh_kinc	0.6596
formula_dale	0.6565
formula_coleman	0.5063

Таблица 2. Степень корреляции признаков на основе формул читабельности

Признак	Коэффициент корреляции
им	-0.5163
страд	0.4816
прич	0.4712
PR	0.4659
сред	0.4534
род	0.3836
SPRO	-0.3722
твор	0.3712
действ	0.3488
ед	-0.3391

Таблица 3. Корреляция грамматических признаков

Как мы видим из Таблицы 1, почти все признаки показали достаточно высокую корреляцию со сложностью, при этом лучшими оказались метрики предложения, средние значения показали бóльшую корреляцию, нежели медианные. Все коэффициенты положительные, т.е. с ростом признака растет и сложность.

Формулы читабельности (Таблица 2) показали стабильно высокую положительную зависимость от уровня сложности текста. Лучший результат показала формула SMOG, которая рассчитывается на основе количества слов длиной более N слогов и количества предложений. Стоит заметить, что существенно коэффициент отличается лишь у формулы Колман-Лиау, остальные показатели очень близки между собой, что при необходимости позволяет нам оставить лишь одну из этих формул.

Грамматические признаки (Таблица 3) также показывают корреляцию со сложностью, однако по сравнению с первыми двумя классами, меньшую. Видно, что в топ-10 попали 3 признака, связанные с причастиями, что подтверждает наше предположение о связи сложности с причастными/деепричастными конструкциями. Доля признака PR(предлоги) вероятнее всего указывает на синтаксическую сложность предложения. Интересна отрицательная корреляция доли именительного падежа в тексте: скорее всего, это связано с

бóльшим количество косвенных падежей, а, значит, распространением предложения. Ожидаемо вносит вклад в сложность доля косвенных падежей. Лидирует тут родительный падеж, который используется для выражения партитивных отношений, при согласовании с числами т.д. Здесь стоит отметить, что в подобном исследовании для текстов русского языка как родного [7] также наблюдается лидерство родительного среди других падежей. Еще одной интересной особенностью стало то, что все 10 лучших вариантов считают долю признака в тексте, т.е. наша гипотеза о роли доли некоторых грамматических категорий в предложении (им. падеж, количество глаголов в финитной форме, существительных на предложение) не подтвердилась.

Признак	Коэффициент корреляции
lex_abstract	0.5243
lex_substance	0.5221
lex_physical	-0.1823
lex_virtual	-0.0711

Таблица 4. Корреляция семантических признаков

Вполне ожидаемо наблюдается корреляция со сложностью доли абстрактной лексики и субстанций. Неожиданно мал вклад физических объектов (гипотеза состояла в том, что чем больше конкретных материальных предметов - тем проще текст).

Признак	Коэффициент корреляции
inA2	-0.8181
inB1	-0.7972
inA1	-0.7949
inB2	-0.7175
infr3000	-0.6452
infr5000	-0.6352
infr10000	-0.6148
infr1000	-0.5755
infr500	-0.5459
infr300	-0.5281

Таблица 5. Корреляция лексических признаков.

Абсолютными рекордсменами по связи со сложностью стали лексические минимумы для всех уровней и списки слов с "серединной" частотой от 300 до 10000 слов. Все эти признаки имеют отрицательную корреляцию, что означает чем больше слов в тесте покрываются заданным списком слов, тем он проще.

Таким образом были отобраны лучшие признаки с коэффициентом корреляции выше 0.3). Из 129 признаков таких оказалось 44. Лидирующие позиции в этом списке заняли лексические признаки (лексические минимумы и частотные списки слов), что подтверждает нашу теорию о большом влиянии лексики на сложность иностранного текста. Среди частотных списков наиболее информативными оказались «медианные» списки, от 300 до 10000 слов, слишком маленькие и слишком большие оказались не так эффективны. Большую корреляцию также показали формулы читабельности. Грамматические признаки также доказали свой вклад в понятие сложности текста, они составили 17 из 44, хотя их коэффициенты оказались и не так велики.

Стоит отметить здесь некоторые особенности полученных коллекций признаков, которые необходимо учесть в дальнейшей работе. Во-первых, очевидно, что полученные признаки во многом коррелируют между собой: частотные списки, формулы читабельности, традиционные метрики текста. Учитывая этот факт, для построения предсказательной модели нами в дальнейшем была использована гребневая регрессия (Ridge Regression), справляющаяся с проблемой мультиколлинеарности признаков.

Второй особенностью является возможная нелинейность грамматических признаков. Так, на Рисунке 1 показан график распределения доли деепричастий в тексте, при этом показаны средние и максимальные значения. По графику становится ясно, что до 3 уровня студенты вообще не знают деепричастия, далее они изучают эту тему и по всей вероятности активно тренируют её, отсюда берется скачок максимального значения, далее график выравнивается. В целом мы все же видим рост доли в зависимости от уровня, но подобные факты заставляют нас осторожнее относится к выбору модели машинного обучения.

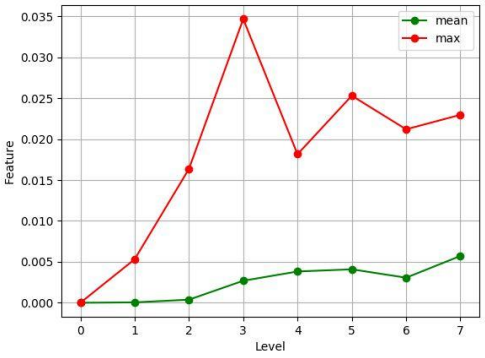

Рисунок 1. Распределение средних и максимальных значений доли деепричастий в тексте

Заключение и направления дальнейшей работы

Полученные результаты доказывают существенную корреляцию признаков на основе лексических минимумов и частотных списков и формул читабельности со сложностью русских текстов как иностранных. Среди семантических признаков стоит выделить списки абстрактных слов и субстанций. Некоторые морфологические признаки также показали связь со сложностью - лидерами этой категории признаков стали именительный и родительный падежи, средний род, количество предлогов, а также признаки, связанные с долей причастий и деепричастий в тексте.

В качестве основных направлений дальнейшей работы мы рассматриваем расширение корпуса, подключение синтаксической и словообразовательной информации, дальнейшая отладка методов машинного обучения. Также интересно было бы сравнить вклад вышеописанных признаков в сложность текстов на других языках.

Библиография.

1. Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Proceedings of HLT-NAACL’07. 460–467.
2. K. Collins-Thompson. Computational assessment of text readability: a survey of current and future research. In: François, Thomas and Delphine Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics, 2014, pp. 97-135
3. Nasser Zalmout, Hind Saddiki and Nizar Habash. Analysis of Foreign Language Teaching Methods: An Automatic Readability Approach. Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016), 2016.
4. Serge Sharoff, Svitlana Kurella, and Anthony Hartley. Seeking needles in the web’s haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8), 2008.
5. Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How Accurate are Lexile Text Measures? Journal of Applied Measurement. 7(3), 2006, pp 307-322.

6. Wright, B. D. & Stenner, A. J. Readability and Reading Ability. Paper presented to the Australian Council on Education Research, 1998.
7. Дружкин К.Ю. Метрики удобочитаемости для русского языка. Выпускная квалификационная работа, Национальный исследовательский университет «Высшая школа экономики», Москва, 2016.
8. О. Н. Ляшевская, С. А. Шаров. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.
9. Оборнева, И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров : диссертация ... к.п.н.: Москва, 2006.
10. Томина Ю.А . Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство) : диссертация ...к.п.н.: Москва, 1985.