А.Б. Кутузов
akutuzov72@gmail.com
ТюмГУ
Опубликовано в:
Вестник Нижегородского государственного лингвистического университета им. Н.А. Добролюбова. Вып.4. Лингвистика и межкультурная коммуникация. - Нижний Новгород: НГЛУ, 2009 г.

Методики определения сложности текста в рамках переводческого анализа

Abstract

The paper overviews the existing ways to calculate complexity of original text parsing by human translator or by machine translation system. The author emphasizes two algorithms: hierarchy of text predicates and approximate calculation of text information entropy through compressing it with mainstream software like gzip. Preliminary verification of the last algorithm on the corpus of training texts for translation is described.

Аннотация

Работа посвящена обзору существующих методик определения сложности обработки текста оригинала человеком-переводчиком или системой машинного перевода. Автор выделяет как наиболее важные алгоритм иерархии текстовых предикатов и алгоритм приблизительного вычисления информационной энтропии текста через его сжатие существующими программами компрессии. Описана пилотная проверка последнего алгоритма на материале учебных текстов для перевода.

Введение

Само появление переводческих ошибок обусловлено двумя факторами: компетенцией самого переводчика и сложностью переводимого текста. Существует явная потребность в определении сложности текста, подлежащего переводу. Это необходимо как практикующим переводчикам для оценки стоимости перевода и затрат времени, так и преподавателям для определения уровня текста и его соответствия этапам обучения. Кроме того, это нужно и для систем автоматического перевода (machine translation).

Любой перевод происходит в два этапа: обработка (parsing) текста оригинала и порождение (синтез) текста перевода. Мы рассматриваем определение сложности или трудности обработки текста оригинала вне зависимости от его языка.

Встаёт вопрос, каким же вообще закономерностям подчиняется определение сложности текста с точки зрения переводчика? Что такое эта «**сложность текста**»? Интуитивно понятно, что этот параметр прямо пропорционален количеству тех лексических единиц, синтаксических структур и прагматических ситуаций, которые вызывают трудности в переводе [1]. Но эти факторы нестабильны — ведь для каждого переводчика большую трудность представляет свой набор лингвистических единиц. Так, текст с большим количество юридических терминов будет воспринят как сложный переводчиком художественной литературы и как простой — переводчиком, специализирующимся в области

юриспруденции. Но проблема не столько в этом, а в том, что вряд ли возможно даже гипотетически представить себе алгоритм определения сложности текста на основе этих факторов, который бы подходил к любому тексту и не был субъективным. Возможно, именно поэтому литературы по проблеме определения сложности перевода очень мало.

В данной работе мы описываем существующие объективные алгоритмы определения сложности любого текста. Многие из этих алгоритмов создавались без учёта переводческой специфики. Тем не менее, автор надеется на то, что этот материал будет полезен переводчикам и переводоведам. Интересно, что фактически каждый из представленных ниже алгоритмов предполагает собственное понимание самой концепции «сложность текста».

Методы и алгоритмы

1) Повторяемость

Для переводчика наиболее очевидный практический способ определить сложность текста — это вычислить его внутреннюю повторяемость и повторяемость по сравнению с текстами, переведёнными раннее. Естественно, что если в представленной к переводу новой версии документа по сравнению со старой версией изменилось лишь несколько строк, то сложность текста резко падает. То же самое происходит, если текст обладает высокой степенью внутренней повторяемости хотя бы на уровне предложения. Строго говоря, это вообще не оценка сложности текста, но в практическом плане такой метод имеет право на существование.

Для оценки повторяемости и распространения (propagation) уже переведённых сегментов на новый документ служат системы памяти переводов (translation memory), например, OmegaT или Trados.

В данном случае сложность текста понимается в сугубо практическом плане - как количество уникальных единиц, не переводившихся ранее.

2) Специализация

Ещё один алгоритм определения сложности текста — отношение количества терминов к общему количеству слов. **Гипотеза** в данном случае такова: чем больше в тексте терминов, тем сложнее он для перевода. Однако, у этого алгоритма есть две проблемы — во-первых, не очевидна истинность гипотезы (возможно, что однозначность и системность терминов наоборот способствует лёгкости перевода), а во-вторых, до сих пор остаётся не до конца решённым вопрос о том, что же, собственно, считать терминами. Тем не менее, в рамках одного тематического поля вполне можно реализовать алгоритм обработки текста, основанный на одном или нескольких специальных словарях. Такой алгоритм мог бы оценивать количество словарных терминов в данном тексте и возвращать оценку отношения этого количества к общему объёму текста.

Итак, данный подход под сложностью текста понимает степень его специализированности.

3) Синтактика

Попытки определения сложности парсинга текста предпринимались в психолингвистике. Эта наука рассматривает обработку текста как поиск соответствия его информационной структуры имеющимся в памяти человека фреймам ситуаций. Для

выявления трудностей этого процесса в психолингвистике обычно используется так называемый предикатный подход [3], который имеет дело с текстовым субъектом, и иерархией текстовых предикатов, которая может быть представлена в виде дерева. **Гипотеза** звучит так: чем сложнее дерево предикатной структуры (синтаксис текста) — тем сложнее парсинг текста. Поскольку задача автоматического синтаксического анализа для основных языков в принципе уже решена [7], то можно использовать любой синтаксический модуль (например, *Synan* рабочей группы aot.ru) и затем определённым образом параметризовать его вывод (например, подсчитывать число синтаксических связей и присваивать им весовые категории).

Тем не менее, такая задача требует наличия специализированных программ и разработки методов параметризации их вывода. Пока это довольно трудоёмко и вряд ли пригодно для повседневного использования. Однако мы считаем данный метод объективным и весьма перспективным в будущем.

4) Соотношение types/tokens и средние длины слов и предложений

Ещё один простой способ определения сложности текста пришёл, как ни странно, из генетики. Там используют лингвистические методы при обработке данных о ДНК [2]. В рамках этого подхода сложность текста есть функция от богатства его словаря (лексического разнообразия). Используется следующая **гипотеза**: чем больше в тексте словоформ (word types) по отношению к словоупотреблениям (word tokens), тем сложнее текст. Самым сложным текстом является тот, в котором количество словоупотреблений равно количеству словоформ. Если же текст лексически беден, то его обработка не сложна. Существует множество программных модулей, реализующих автоматический подсчёт соотношения словоформ и словоупотреблений.

Вариация этого алгоритма — оценка отношения фактического количества слов в тексте к их максимально возможному количеству (то есть, к количеству букв). Как легко видеть, здесь оценивается длина слов. Собственно, эта методика известна ещё с 50-х годов под названием «флэш-тест» (употребляется в основном в рекламном деле). **Гипотеза**: сложность текста прямо пропорциональна средней длине слова и средней длине предложения.

Эти алгоритмы просты в употреблении, но, к сожалению, не отражают взаимозависимости между единицами перевода, которые содержатся в тексте. Текст с большим количеством длинных, но похожих друг на друга слов может быть легче для парсинга, чем текст с короткими, но очень несхожими лексическими единицами.

Мы считаем, что подобные методики не определяют напрямую сложность текста, но могут послужить одним из этапов такой оценки. Речь идёт об автоматическом определении функционального стиля текста по спектрам длин слов (а иногда и просто через среднюю длину слова). Такие компьютерные классификаторы уже существуют (см., например, [8]) и вполне способны различить разговорную речь, научные и деловые статьи, публицистику, новости и художественную прозу [6]. Подобный предварительный разбор может быть полезен и в рамках предпереводческого анализа. Отметим, что надёжное определение стиля начинается на размерах текста от 10 тысяч слов.

5) Информационная энтропия

Наконец, последний алгоритм определения сложности текста — чисто математический. Он основан на понятии **информационной энтропии** (напомним, что это мера хаотичности информации). Общий смысл этого алгоритма — вычисление меры

избыточности или предсказуемости текста. **Гипотеза** здесь звучит так: чем менее предсказуем и избыточен текст, тем он сложнее.

Встаёт вопрос — как же быстро и объективно определить сложность текста, основанную на энтропии? В теоретическом плане ответ на это дал известный математик А.Н. Колмогоров [4]. В его терминах, сложность текста — это длина минимальной программы, которая выводит данный текст, а энтропия — это сложность, делённая на длину текста.

К сожалению, это определение чисто умозрительное. Надёжного способа однозначно определить эту программу не существует.

Но есть алгоритмы, которые фактически как раз и пытаются вычислить колмогоровские сложность текста и энтропию [5]. Что интересно, эти алгоритмы знакомы всем, кто работает с компьютером. Это так называемые «архиваторы» или «компрессоры» - программы, сжимающие файлы. В самом деле, текст, сжатый программой zip или rar, представляет собой некоторую «программу», которая затем при декомпрессии интерпретируется таким образом, что на выходе мы видим исходный текст. Конечно, это не идеальная колмогоровская «минимальная программа», но некоторое приближение к ней.

Таким образом, по степени сжатия текста какой-либо программой компрессии, мы можем судить о его избыточности и в какой-то степени о мере его сложности. Если текст сжимается хорошо (то есть, получаем сжатый файл меньшего размера), следовательно, он обладает высокой избыточностью, а сложность его не велика. Если же сжатый текст по размеру почти не уступает несжатому, то мы можем сделать заключение о его относительно высокой сложности. В пользу данного метода говорит его быстрота и доступность — практически на любом компьютере есть программа-архиватор.

В следующем разделе мы проверим, действительно ли колмогоровская энтропия текста коррелирует с переводческими задачами.

Результаты

Для предварительной проверки корреляции мы взяли три текста на английском языке из рабочих программ кафедры перевода и переводоведения ТюмГУ для 3, 4 и 5 курсов, соответственно. Наша гипотеза состояла том, что тексты для старших курсов должны быть сложнее, чем для младших, а следовательно — обладать более низкой избыточностью и большей энтропией.

Мы сжали каждый из текстов компрессором gzip 1.3.5. Он реализует алгоритм сжатия Лемпеля-Зива (LZ77). Использовались настройки сжатия по умолчанию.

Вот полученные оценочные результаты:

[kender@neverland test]$ gzip -l *

compressed	uncompressed	**ratio**	uncompressed_name
1226	2505	**52.0%**	3.txt
762	1347	**45.2%**	4.txt
1197	2221	**47.2%**	5.txt

Больше всего нас интересует столбец ratio, то есть, степень сжатия. Чем она выше, тем меньше избыточность исходного текста, тем меньше в нём энтропии. Как можно видеть, действительно, текст для третьего курса сжимается гораздо лучше, чем тексты для 4 и 5

курса. Различия между их степенями сжатия (45,2 и 47,2 процента), скорее всего, объясняются статистической погрешностью.

Конечно, нашей выборки недостаточно для каких-либо окончательных выводов, это всего лишь пилотный «забор проб». Тем не менее, проверка показала, что определённая корреляция есть, метод определения сложности текста через его информационную энтропию показывает осмысленные результаты, а следовательно — можно проводить дальнейшие исследования в этой области.

Заключение

Итак, мы показали некоторые из существующих концепций сложности текста и алгоритмы её вычисления: сравнение с раннее переведёнными текстами, терминологический, психолингвистический, алгоритм спектров длин слов / богатства словаря и метод информационной энтропии. Все они могут быть использованы для определения того или иного аспекта сложности парсинга исходного текста. Из перечисленных алгоритмов наиболее перспективным в рамках переводоведения нам представляется **психолингвистический** (оценка синтаксической сложности через предикатный подход), а наиболее готовым к применению прямо сейчас — **энтропийный**. Возможно, в ближайшем будущем мы увидим какие-то варианты синтеза этих двух подходов.

Литература

1. **Helge Dyvik**. The Interaction Between Text Difficulty and Translation Accuracy // Out of Corpora: Studies in Honour of Stig Johansson.- Rodopi, 1999, ISBN 904200505X
2. **Olga Troyanskaya et al**. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. // Bioinformatics, volume 18, number 5, may 2002 — pp. 679-688
3. **Данилова Н.Н.** Психофизиология:Учебник для вузов. –М.: Аспект-Пресс.1998.-373 с.
4. **Колмогоров А.Н.** Три подхода к определению понятия «количество информации»//Пробл. передачи информ. 1965. Т.1. N0 1, С.3-11.
5. **Кукушкина О.В., Поликарпов А.А., Хмелёв Д.В**. Определение авторства текста с использованием буквенной и грамматической информации//Проблемы передачи информации, 2001, т.37, вып.2, с.96-108. Translated in "Problems of Information Transmission", pp. 172-184. URL: http://www.philol.msu.ru/~lex/khmelev/published/gramcodes/gramcodeswin.html
6. **Хмелёв Д.В**. О лингвоанализаторе 3-эпсилон. URL: http://www.philol.msu.ru/~lex/khmelev/descrwin.html
7. Рабочая группа «Автоматический анализ текста», URL: http://www.aot.ru
8. «Худломер», URL: http://teneta.rinet.ru/hudlomer/