УДК 004.912

ПРОБЛЕМЫ ПЕРВИЧНОГО АНАЛИЗА ЕСТЕСТВЕННЫХ ТЕКСТОВ

А. Н. Владимиров

Воронежский государственный университет

Поступила в редакцию 12.07.2013 г.

Аннотация. В данной статье описаны основные проблемы, с которыми сталкиваются анализаторы естественного теста на этапе первичного анализа. Приведены примеры случайных и преднамеренных искажений текста.
Ключевые слова: анализ текста, опечатка, естественный язык, парсер.
Annotation. This topic lists the most common problems of natural language pre-semantic analysis and gives some examples of mistypes and deliberate distortions of natural text.
Keywords: text analysis, mistype, natural language, parser.

Развитие информационных технологий и проникновение интернета в жизнь каждого человека делает возможным обильное общение и обмен информацией в сети и порождает огромное количество текста, преимущественно на естественном языке. А уже обилие такой информации представляет собой интересный объект для изучения и анализа [1–5].

На сегодняшний день существуют ситуации, когда без анализа текста просто не обойтись – например, родительский контроль, ну или интеллектуальное построение фразовых индексов по различным форумам, рассылкам, и прочим местам массового общения людей.

Типичный порядок обработки текста включает в себя 2 пункта – первичный (предсемантический) и семантический анализ. В свою очередь, первичный анализ разбивается на лексический (выделение лексем – отдельных слов), морфологический (выделение морфем – базовой словоформы и формы слова) и синтак сический (определение связей между отдельными морфемами). Между тем, анализ текста на естественном языке – не такой простой процесс, как может показаться на первый взгляд. При анализе естественного языка зачастую возникают некоторые проблемы.

СЛОЖНОСТЬ ФОРМАЛИЗАЦИИ

Естественный язык редко удается четко формализовать, в отличие, например, от языков программирования, в которых всегда существуют некоторые правила. Например, порядок слов. В тексте, написанном на неком языке программирования, всегда известно, что оператор следует за открывающей скобкой, а за ним, в свою очередь, идет либо следующий оператор, либо скобка закрывающая. В естественном языке есть прямой и обратный порядок слов, безличные предложения, разные сокращенные формы, подразумевающиеся члены предложения, поэтому предположить какой член предложения будет следующим, крайне сложно.

ПРОБЛЕМА ДИАЛЕКТОВ

Если даже взять язык с достаточно простыми правилами, например испанский, то об этих правилах можно говорить лишь в рамках определенного диалекта. Для испанского языка основным является кастильский диалект, других диалектов с учетом распространенности по всему земному шару, можно насчитать много десятков, если не сотен. Большинство отличий в этих диалектах – в произношении слов. В испанском языке есть буква ‘ll’, которая в зависимости от региона может читаться как нечто среднее между ‘й’ и ‘л’ в каталонском, и как ‘ж’ в некоторых латиноамериканских странах. Также могут иметься отличия в лексике, к примеру банан в разных испаноговорящих странах может называться banano, bananero, plбtano. Отличия могут быть в употреблении местоимений, предлогов, глагольных форм (синтаксические) и образовании слов (морфологические).

Из вышеперечисленных отличий первые 2 не оказывают значительного влияния на анализ текста, они могут оказывать влияние лишь на

содержание словарей, в то время как синтаксические и морфологические отличия могут сильно влиять на подход к анализу различных диалектов одного языка.

Ошибки и опечатки

Первичный анализ текста часто используется в местах, где происходит общение людей – различные интернет-форумы, новостные группы, конференции. И далеко не все люди, которые общаются в сети, следуют при написании сообщений правилам языка, на котором они общаются. Встречаются как ошибки связанные с неграмотностью отдельных субъектов общения, так и ошибки, связанные с местом этого общения (какие-то особенности общения на определенном форуме) либо связанные с тем, что приходится печатать много и быстро, то есть опечатки.

Если с проблемами, связанными с заведомо правильным текстом, можно справиться достаточно безболезненно, то с текстом, написанным с опечатками, это становится намного сложнее. Типичными опечатками можно считать следующие:

1. Пропуск, замена или добавление лишнего символа (слово – слво – слоло – словоо).

2. Перестановка двух букв местами (буква – бкува).

3. Лишний или отсутствующий пробел (красный свет – красн ый свет – красный-свет).

4. Печать со смещением на клавишу (привет – апмыки).

5. Печать в другой раскладке (раскладка – hfcrkflrf).

6. Печать соседней буквы по клавиатуре (клавиатура – коавитурв).

Здесь надо понимать, что учет наличия опечаток при анализе ведет к увеличению количества времени, необходимого для обработки текста. Если каждое слово проверять на наличие произвольной опечатки в каждой позиции, происходит комбинаторный взрыв из-за огромного количества вариантов исправления.

Особенно следует отметить опечатки, представляющие собой пропуск и добавление лишнего пробела. Типичная схема работы системы анализа текста предполагает проведение лексического анализа до того, как будет работать морфологический, синтаксический и семантический анализатор, то есть слова выделяются из предложения до того, как можно говорить о каком-либо поиске слова в словаре, построении синтаксического графа и проверке смысла всего предложения. Однако в случае с добавлением или пропуском пробелов такая схема не будет правильно идентифицировать подобные опечатки, приводя к противоречиям на уровне морфологического, синтаксического или семантического анализа. Система просто идентифицирует одно слово как два или, наоборот, два как одно, и дальше могут возникать разные варианты поведения:

1. Если слово разбилось(соединилось) так, что выделенная лексема не может быть найдена в словаре, система либо выдаст ошибку о том, что не смогла найти слово, либо попробует найти слово, из которого при опечатке получается текущее. Например, разбиение пополам слова «семантический» на «семант» и «ический» вряд ли будет опознано с использованием словаря, в то время как разбиение слова «искажение» на «иска» и «жение» может быть воспринято как опечатки в словах «искра» и «жжение». Слово может разбиться или соединиться так, что получится словарное слово в чистом виде, например слово «разбилось» отлично расщепляется на «разбил» и «ось».

2. Если система всё же разбила предложения на слова (применяя исправления или нет), то на этапе синтаксического анализа могут появиться неожиданные несогласованные члены предложения. Например, если взять предложение «Слово разбилось на два», и сделать а нем опечатку «Слово разбил ось на два», то на этапе синтаксического анализа возникнет 2 противоречия – несогласованность рода подлежащего и сказуемого, а также рода подразумеваемого дополнения и числительного, относящегося к нему.

3. Если система всё же успешно завершила этап синтаксического анализа и построила некую структуру предложения, то смысл предложения будет неправильно понят, и если система предполагает зависимость смысла от контекста, то неправильная трактовка смысла одного предложения может привести к неправильной трактовке соседних.

Таким образом, возникновение опечатки в виде одного-единственного лишнего или пропущенного пробела при отсутствии тщательной обработки подобной ситуации может приводить к искажению всего смысла текста. Причем, как и в случае с любыми опечатками, решение про-

блемы перебором приводит к комбинаторному взрыву, тем более что при опечатках в рамках одного слова дальнейший анализ не требует дополнительного лексического анализа, а в случае с пробелами может потребоваться возврат на этап лексического разбиения и с момента морфологического анализа, и из синтаксического анализатора, что накладывает на систему определенные требования к взаимодействию программных модулей.

НАМЕРЕННОЕ ИСКАЖЕНИЕ

Помимо случайных ошибок, возникающих вследствие невнимательности или опечаток, иногда встречаются и ошибки, сделанные преднамеренно. Обычно это связано с попыткой обойти защиту различных сетевых ресурсов от нецензурной лексики, преодолением родительских фильтров, а также предложением товаров и услуг. Отдельно можно выделить попытки рассылки нежелательной корреспонденции (спама), нацеленные на преодоление различных фильтров на уровне почтовых серверов.

Типичными вариантами намеренного искажения можно считать:

1. Замена буквы идентичной по написанию латинской буквой (привет – npuBeT).

2. Замена сочетания букв цифрой – (информация – ин4мация).

3. Замена глухой согласной на звонкую и наоборот (лишний – лижний).

4. Транслитерация слова в латинскую раскладку (предложение – predlojenie).

5. Замена схожей по звучанию латинской буквой (словосочетание – sловосochетание).

6. Замена буквы сочетанием букв, сходных по звучанию, и наоборот (счастье – щастье).

Исправление намеренных ошибок представляется еще более сложным и ресурсоемким, нежели исправление случайных опечаток, так как намеренное искажение направлено именно на затруднение работы текстовых анализаторов. Как уже было отмечено ранее, при исправлении различных ошибок может критически снижаться скорость текстового анализатора из-за огромного числа вариантов перебора. Решение подобной задачи, что называется «в лоб», видится маловозможным с использованием современных вычислительных ресурсов. Поэтому для исправления ошибок могут применяться различные методики – расширение словарей; вероятностные алгоритмы, связанные с частотой встречаемости определенных ошибок; объединение этапов первичного анализа; построение специальных индексов для ускорения поиска по словарю. Только сочетая несколько подходов, комбинируя алгоритмы и структуры, можно действительно подойти к решению задачи синтаксического произвольного текста.

СПИСОК ЛИТЕРАТУРЫ

1. *Селезнев К.Е.* Лингвистика и обработка текстов / К.Е. Селезнев, А.Н. Владимиров // Открытые системы. – 2013. – № 4. С. 46–49.

2. *Артемов М.А.* Оптимальная организация морфологических словарей в поисковых системах / М. А. Артемов, К.Е. Селезнев, А.П. Якубенко // Вестник ВГУ Сер.: Системный анализ и информационные технологии. – 2006. – № 2. – С. 156–161.

3. *Селезнев К.Е.* Обработка текстов на естественном языке / К.Е. Селезнев // Открытые системы. – 2003. – № 12. С. 48–53.

4. *Артемов М.А.* Анализ совстречаемости слов / М.А. Артемов, К.Е. Селезнев, В.А. Сорокина // Вестник Воронеж. гос. ун-та. Сер. : Системный анализ и информационные технологии. – Воронеж, 2012. – № 2. – С. 159–163.

5. *Толдова С.Ю.* Оценка методов автоматического анализа текста 2011 –2012: синтаксические парсеры русского языка. / С.Ю. Толдова [и др.]. – URL : http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Toldova.pdf (дата обращения: 03.07.2013).

Владимиров Александр Николаевич – аспирант, Воронежский государственный университет, факультет прикладной математики информатики и механики, кафедра программного обеспечения и администрирования информационных систем. E-mail: alcobass@gmail.com

Vladimirov Alexander – postgraduate, Voronezh State University, Applied Mathematics, Informatics and Mechanics department. E-mail: alcobass@gmail.com