УДК 004.421.6

Р.М. Бабаков, А.Д. Леонов
Донецкий национальный технический университет, г. Донецк
кафедра систем искусственного интеллекта

МЕТОДЫ АВТОМАТИЗИРОВАННОЙ КОРРЕКЦИИ СПЕЦИАЛИЗИРОВАННЫХ ПРИРОДНО-ЯЗЫКОВЫХ ТЕКСТОВ

Аннотация

Бабаков Р.М., Леонов А.Д. Методы автоматизированной коррекции специализированных природно-языковых текстов. *Выполнен анализ методов обработки текстовой информации. Произведен анализ методов автоматизированного выявления орфографических ошибок в текстах. Выбран оптимальный метод для экономии трудовой деятельности человека и минимального пропуска текстовых ошибок.*

Ключевые слова: *методы обработки текстовой информации, методы автоматизированного выявления орфографических ошибок.*

Постановка проблемы. Большое количество пользователей ПК часто сталкиваются с проблемой огромного количества орфографических и пунктуационных ошибок при написании курсовых, диссертаций и различных документов. Эффективным решением этой важной и повседневной проблемы является автоматизация процессов обработки больших объемов текстовой информации. Таким образом, возникает необходимость разработки системы автоматизированной коррекции специализированных природно-языковых текстов. Для её реализации необходимо выполнить следующие этапы:

– изучить распространенные методы обработки текстовой информации;

– выявить основные шаги автоматизации процесса исправления текста.

Цель статьи – провести анализ методов автоматизированной коррекции специализированных природно-языковых текстов и выбрать оптимальный метод для экономии трудовой деятельности человека и минимального пропуска текстовых ошибок.

Анализ методов обработки текстовой информации. Обработка информации – это процесс преобразования исходной информации из

одного вида в другой, осуществляемое по строгим формальным, заданными заранее, правилами.

Существует два распространенных обработки текстовой информации:

- автоматическая классификация;
- автоматическое обнаружение орфографических ошибок.

Выявление орфографических ошибок является одним из наиболее трудоемких и дорогостоящих процессов обработки информации [1]. Наряду с неуклонным ростом объемов электронной научно-технической информации, растет и острота этой проблемы. Поэтому в последние годы увеличился интерес к данной проблеме. В настоящее время разработано множество различных методов, которые позволяют частично или полностью решать эту проблему. Программы автоматического обнаружения и исправления ошибок в текстах на естественных языках (назовем их автокорректор – АК) получают все большее распространение. Говоря точнее, АК производят автоматически только выявление ошибок и то лишь орфографического типа, а, собственно, коррекция ведется обычно с участием человека. Но и при таком ограниченном автоматизме точность и производительность проверки текстов с помощью АК повышаются настолько значительно, что АК становятся неотъемлемой частью процесса подготовки различных деловых документов, статей, рефератов, книг.

Методы выявления ошибок. Известно, по крайней мере, три метода автоматизированного выявления орфографических ошибок в текстах [2]:

- статистический;
- полиграммный;
- словарный.

При статистическом методе из текста одна за другой выделяются составляющие его словоформы, а их перечень по ходу проверки упорядочивается согласно частоте. По завершении просмотра текста упорядоченный перечень предъявляется человеку для контроля, например, через экран дисплея. Орографические ошибки и опечатки в сколько-нибудь грамотном тексте несистематические и редкие, поэтому такие слова оказываются где-то в конце списка. Заметив их, контролирующее лицо может автоматизировано найти их в тексте и исправить.

При полиграммном методе все двух- или трехбуквенные сочетания, которые встретились в тексте (биграммы и триграммы) проверяются по таблице их допустимости в данном естественном языке. Если в

словоформе не содержится недопустимых полиграмм, она считается правильной, а иначе - сомнительной, и тогда предъявляется человеку для визуального контроля и, если нужно, исправления.

При словарном методе все входящие в текст словоформы, после упорядочивания или без него, в своем первоначальном текстовом виде или после морфологического анализа, сравниваются с содержанием заранее составленного машинного словаря. Если словарь такую словоформу допускает, она считается правильной, а иначе предъявляется контроллеру. Он может оставить слово как есть; оставить его и вставить в словарь, так что далее в сеансе подобное слово будет распознаваться системой без замечаний; заменить (исправить) слово в данном месте; потребовать подобных замен по всему дальнейшему тексту; отредактировать слово вместе с его окружением. Операции над сомнительным участком текста, указанные или другие возможные, могут комбинироваться исходя из замысла проектировщика АК.

Результаты неоднократных исследований показали, что только словарный метод экономит труд человека и ведет к минимуму ошибочных действий обоих родов - пропуска текстовых ошибок, с одной стороны, и отнесения правильных слов к сомнительным, с другой [2]. Поэтому словарный метод стал доминирующим, хотя полиграммный метод иногда и применяют как вспомогательный.

Автоматизация процесса исправления. Можно предложить три степени автоматизации процесса коррекции текста:

- только выявление ошибок;
- выявление их и выдвижение гипотез (альтернатив, кандидатов) по исправлению;
- выявление ошибок, выдвижения гипотез и принятие одной из них (если хотя бы одна выдвинута системой) как автоматически внесенного исправления.

Без первой степени АК немыслим.

Вторая и третья степень возможны только при словарном методе. Уже второй метод существенно облегчает внесение исправлений, потому что в большинстве случаев исключает перенабор сомнительного слова. Особенно полезны найденные альтернативы, когда контролирующее текст лицо нетвердо знает этот естественный язык или конкретную терминологическую область. Однако выдвижение гипотез требует больших переборов с поиском в словаре. Поэтому современные АК часто имеют средство выдвижения гипотез только как факультативного, запускаемого если нужно, избирательно для данного сомнительного слова.

Третья степень автоматизации привлекательная и одновременно опасная. Прелесть заключается в полной автоматизации процесса исправления. Опасность же в том, что ни один словарь, в том числе заключенный в человеческом мозге, никогда не бывает исчерпывающе полным. Когда незнакомое слово встречает система, основанная на неполном словаре, она может «исправить» его на ближайшее ей знакомое, однако резко изменив исходный смысл текста. Особенно опасно править имена лиц, фирм, изделий.

Чисто автоматическому исправлению мог бы способствовать автоматический синтаксический и семантический анализ проверки текста, но он еще не стал принадлежностью обычных АК. И даже при его наличии только человек сможет диагностировать быстро меняющиеся совокупности имен, терминов и аббревиатур, а также окказионализмы – словесные новации, появляющиеся случайно.

В связи со сказанным, полная автоматизация исправлений может применяться только в любом из следующих ограничительных условий:

– текст выглядит как перечень терминов и терминологических словосочетаний в стандартной их форме, так что в АК достаточно иметь словарь по объему и проблематике. При этом все термины между собой «непохожие»;

– ошибки носят характер замены кодов исходных букв на коды букв, совпадающих или близких к исходным по изображению. Например, замена кодов ASCII русских букв А, В, С, Е, У на коды латинских букв; латинские буквы I и 0 – на цифры I и 0 и т.п. Сюда же отнесем повторы одной и той же буквы, что возникают из-за продолженного натиска клавиши или её неисправности. В подавляющем большинстве, если в словоформе более 2-3 букв, такие исправления абсолютно правильные.

Выводы. Выполнен анализ методов обработки текстовой информации и выявления ошибок. Произведен анализ автоматизации процесса исправления орфографических ошибок в текстовой информации. Выбран оптимальный метод для экономии трудовой деятельности человека и минимального пропуска текстовых ошибок.

Список литературы

1. N-граммные методы обработки текстовой информации. [Электронный ресурс]. – Режим доступа: http://gpntb.ru/win/inter-events/crimea95/report/rep075_r.html.
2. Peterson J.L. Computer programs for detection and correction spelling: errors / Commun. ACM, 1980, № 12 – P. 676—687.