Секция «Вычислительная математика и кибернетика»

Метод автоматического исправления ошибок сочетаемости слов в текстах

Азимов Александр Евгеньевич
Аспирант
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: mitradir@gmail.com

Метод автоматического исправления ошибок сочетаемости слов в текстах
Азимов А.Е.
Аспирант
Московский государственный университет имени М.В.Ломоносова,
факультет вычислительной математики и кибернетики, Москва, Россия
факультет вычислительной математики и кибернетики, Москва, Россия
Email: mitradir@gmail.com

Современные системы редактирования текстов на естественном языке, использующие компьютерные словари и методы синтаксического анализа, выявляют орфографические и частично синтаксические ошибки, однако пока не обнаруживают и не корректируют ошибки сочетаемости слов. Подобные лексические ошибки (дать внимание вместо уделить внимание; показывать образец поведения вместо показывать пример поведения) нарушают сложившиеся в конкретном естественном языке правила сочетаемости слов. В то же время ошибки такого рода нередко допускаются людьми даже в родном языке, а в случае иностранного языка они возникают гораздо чаще.

В настоящий момент разрабатываются методы коррекции ошибок сочетаемости в текстах, основанные на использовании статистики (частот) встречаемости слов, например [1], нацеленных на исправления частных видов словосочетаний. В данной работе описывается разработанный метод автоматической коррекции ошибок сочетаемости в текстах естественного языка, полученных в результате перевода с помощью словаря или с использованием машинного перевода. Предполагается, что основной причиной возникновения ошибок сочетаемости является часто используемая стратегия пословного перевода; поэтому ошибки сочетаемости зависят от исходного языка переводчика. Разработанный метод не имеет ограничений на виды исправляемых словосочетаний.

Метод последовательно рассматривает предложения текста: для каждого предложения генерируется набор предложений-замен, наследующих синтаксическую структуру исходного предложения и состоящих из слов, являющихся переводными эквивалентами слов исходного предложения. Для выявления ошибок сочетаемости в предложении была предложена функция, оценивающая соответствие слов предложения его синтаксической структуре и вычисляемая на основе вероятностей синтаксических связей слов. Если значение этой функции для проверяемого предложения меньше значения функции для некоторого предложения-замены, считается, что в исходном предложении допущена ошибка, а данное предложение-замена рассматривается как возможный вариант коррекции.

Описываемый метод был реализован для исправления ошибок сочетаемости в английских текстах, написанных русскоязычными авторами. Для создания базы данных вероятностей синтаксических связей слов был проведен автоматический синтаксический разбор большой коллекции текстов с использованием синтаксического анализатора Stanford Parser [2]. Проведенное тестирование продемонстрировало перспективность предложенного метода: в тестовом наборе было обнаружено 80% ошибок сочетаемости, причем 87% построенных вариантов коррекции предложения включали в себя лингвистически верный вариант исправления.

Литература

1. *Brockett C., Dolan W., Gamon M. Correcting ESL Errors Using Phrasal SMT Techniques // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics ACL , 2006.*

2. *Manning C., Jurafsky D. Stanford Parser 2013 [html] (http://nlp.stanford.edu/software/lex-parser.shtml).*