OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 32.88

МЕТОДЫ ВЕРИФИКАЦИИ ГРАММАТИКИ ЕСТЕСТВЕННЫХ ЯЗЫКОВ ФИННО-УГОРСКОЙ ГРУППЫ НА УРОВНЕ СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

Килеев В.В.[*], Сидоркина И.Г.[*]

[*] *Поволжский государственных технологический университет, г. Йошкар-Ола, Республика Марий Эл, Россия*

slavakileev@yandex.ru

igs592000@mail.ru

В работе приводится обзор методов верификации грамматики естественных языков в контексте семантических представлений для реализации их в компьютерной системе. В работе системы верификации грамматики требуется определять семантику верифицируемого слова для проверки ее соответствия контексту. Отличие методов верификации грамматики естественных языков состоит в способе определения семантики верифицируемого слова. Кроме краткого описания каждого из методов, рассматриваются отличия в получении системой семантической информации каждым из методов.

Ключевые слова: верификация грамматики; получение семантики; множество неоднозначности; обработка естественного языка.

Введение

В области верификации грамматики естественных языков одной из основных целевых задач является выражение смысла (семантики) проверяемого текста. Отличие верификации грамматики от верификации орфографии заключается в том, что во время верификации грамматики проверяется насколько правильно вписывается данное слово в контекст. Таким образом, верификация грамматики осуществляет определение семантики слова и далее проверяет насколько она соответствует семантике всего предложения.

В вопросе верификации текстов естественных финно-угорских языков, проблема верификации орфографии имеет свою реализацию [Килеев и др., 2011]. В то время как проблема определения семантики и верификации грамматики еще только требует своего решения, которое должно учитывать как особенности финно-угорских языков так и особенности построения существующей системы верификации орфографии.

1. Описание проблемы верификации грамматики

Как уже было сказано выше, к вопросам верификации грамматики относится определение правильности употребления слова в контексте. Например, одной из самых частых ошибок является употребление омофонов – слов, произношение которых является одинаковым, но которые пишутся по-разному. Примером омофонов являются слова {quiet, quite} английского языка. Другим примером часто совершаемых ошибок являются слова, которые пишутся почти одинаково, например, {affect, effect}. При этом, в каждом из этих типов ошибок слова образуют множества неоднозначности (confusion sets).

Для правильного принятия решения о правильности употребления того или иного слова в контексте, системе верификации грамматики важно определить смысл каждого слова из множества неоднозначности. Лишь определив это можно сделать правильный выбор. Известно, что все методы верификации грамматики отличаются способами определения семантики каждого слова.

2. Классификация методов

2.1. Общая классификация

Фрагмент общей классификации методов верификации грамматики естественных языков представлен на рис. 1. Ниже в статье представлено краткое описание каждого из методов и дано пояснение об особенностях получения семантики. По возможности также даются пояснения по основным недостаткам и преимуществам методов.

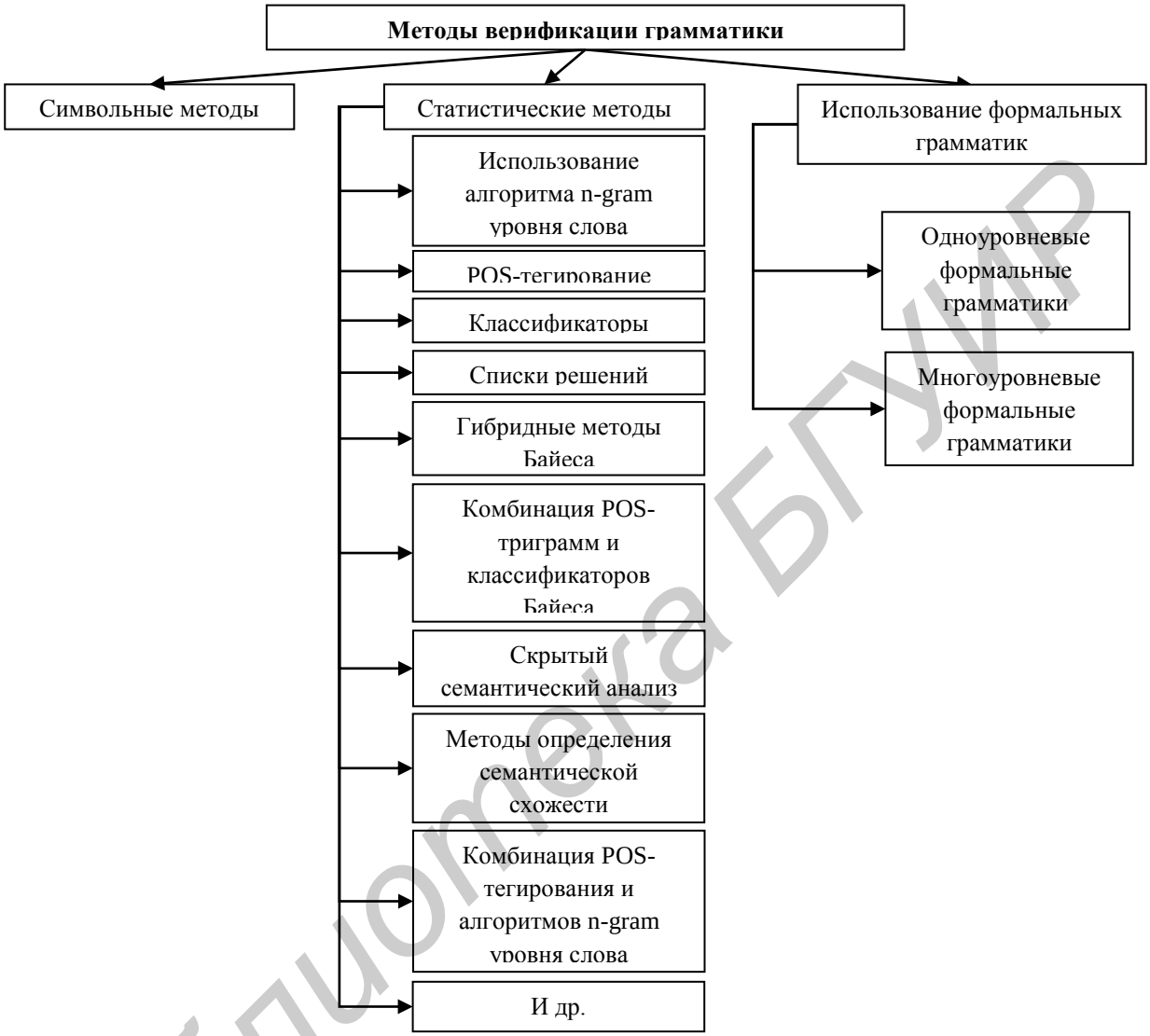

Рис. 1. Фрагмент классификации методов верификации грамматики

2.2. Символьные методы

Суть символьных методов [Naber, 2003] заключается в просматривании всего предложение целиком и поиске грамматических аномалий. Также данные методы иногда называют методами шаблонов. Для каждого варианта совершаемой пользователем ошибки создается шаблон - представление аномалии, на основе которого осуществляется поиск ошибки по тексту. Исправление на правильный вариант также осуществляется на основе шаблона. Каждый шаблон представляет собой заданную «вручную» семантику слов из множеств неоднозначности. Формат записи шаблонов может быть разный,

- O'Hearn, et al.]. N-gram – это последовательность из n следом идущих элементов. В n-gram уровня слова, в качестве таких элементов может использоваться, например, расширяемый язык разметки XML.

2.3. Статистические методы

В данную категорию входит большое количество методов, которые схожи между собой по принципу действия. Всем статистические методы требуют для своей работы наличие лингвистический корпусов. На основе этих корпусов строится база данных, позволяющая разделять семантики разных слов. К статистическим методам относятся следующие:

- Различные реализации n-gram уровня слова [Mays, et al., 1991] [Berlinsky-Schine, 2004] [Wilcox- выступают слова из предложения. Семантика слова w определяется тем, какие слова стоят до и после него в последовательности n-gram. Чаще всего

используются либо биграмм – последовательность из двух элементов, или триграмм – последовательность из трех элементов. Для качественной верификации грамматики, требуется лингвистических корпус. Основной недостаток, что даже при использовании лингвистический корпусов очень больших размеров, правильная триграмм-последовательность может отсутствовать в модели. Классический алгоритм триграмм не учитывает несколько ошибок в предложении, он работает при условии, что в предложении допущена только одна ошибка. Существует также улучшенный алгоритм триграмм, который может учесть несколько исправлений в одном предложении. Вместо выбора одного предложения с максимальной вероятностью, выбираются все предложения с вероятностью выше, чем у текущего. И все изменения применяются последовательно. Если изменения конфликтуют между собой, тогда выбирается то, у которого вероятность выше. Доказано, что данный алгоритм работает лучше, чем проверка грамматики в Microsoft Word 2007 [Hirst, 2008].

- POS тегирование [Marshall, 1983] [Garside, et al., 1987]. У каждого слова в предложении определяется его часть речи, которая берется за основу для представления его семантики.
- Классификаторы Байеса [Gale, et al., 1993]. Основная проблема в том, что для качественной работы данного метода современный лингвистические корпуса слишком малы.
- Списки решений (англ. decision list) [Yarowsky, 1994]. Семантика слова разрешается вручную и определяется присутствием слов-якорей вблизи верифицируемого слова. Строятся таблицы употребляемости слов из множества неоднозначности со всеми возможными словами-якорями с указанием частоты встречаемости получившихся словосочетаний в лингвистическом корпусе. Точность определения семантики слова зависит от размера корпуса.
- Гибридные методы Байеса [Golding, 1995]. Основаны на объединении метода списка решений и классификаторов Байеса. Решение о разрешении семантики строится не на основе одного самого яркого признака, а путем подсчета суммы признаков.
- Комбинация POS-триграмм и классификаторов Байеса, иногда еще называется как Три-Байес (Tribayes) [Golding, et al., 1996]. Семантика вначале определяется при помощи POS-триграмм – триграмм в качестве элементов которого выступают не сами слова о их часть речи, если во множестве неоднозначности стоят слова с одинаковой частью речи, тогда семантика определяется при помощи классификатора Байеса.
- Скрытый семантический анализ (англ. latent semantic analysis) [Jones, et al., 1997]. В этом методе строится матрица в строках которой находятся понятия, а в столбцах – тексты корпуса, на пересечение – значение функции, определяющей частоту появления понятия в тексте. После сингулярного разложения данной матрицы, высчитывается подпространство, содержащее только самые существенные понятия текста. А для получения семантики слова строится его вектор из данного подпространства. За место сингулярного разложения можно также использовать линейный дискриминантный анализ.
- Методы определения семантической схожести [Hirst, et al., 2005] [Budanitsky, et al., 2006]. Все меры семантической схожести строятся на основе специально-размеченного корпуса WordNet.
- Смешанная модель триграмм – это комбинация три-грамм уровня слова и POS три-грамм [Fossati, et al., 2007]. Смешанный триграмм – это последовательность трех элементов (ei, ei+1, ei+2), где элемент ek является либо k-ым словом в предложении, либо его частью речи. При этом накладывается условие, что в последовательности трех элементов, только один элемент может быть словом, остальные должны быть обозначениями частей речи. В остальном данный метод работает также как и классический n-gram.
- Смешанная модель триграмм с добавлением эмпирических законов для построения множества неоднозначности [Fossati, et al., 2008]. За место редакторского расстояния используется вероятности совершения ошибки. Вероятности получают из специального корпуса ошибок (Birkbeck spelling error corpus). Этот корпус содержит пары слово->неправильно написанное слово. Но проблема этого метода заключается в том, что кроме обычного лингвистического корпуса помеченного POS тегами необходим дополнительный корпус совершаемых пользователем ошибок, который даже для английского языка является не полным.

2.4. Методы, использующие формальные грамматики

Данная категория методов разделяется на две большие подкатегории: одноуровневые формальные грамматики [Heidorn, et al., 1986] и многоуровневые формальные грамматики [Bender, et al., 2004]. В одноуровневых формальных грамматиках все грамматические правила записываются в одной грамматике на одном уровне детализации. Многоуровневые формальные грамматики имеют несколько грамматик разного уровня детализации. Логический вывод идет сразу по всем грамматикам. Если в ходе вывода обнаруживается несоответствие грамматик разных уровней детализации, то считается что в этом случае обнаружена ошибка.

3. Выводы

Хотя статистические методы являются одними из самых часто используемых, они требует для своей работы наличие больших лингвистических корпусов. Так как для восточных финно-угорских языков лингвистические корпуса на данный момент отсутствуют, то при разработке системы верификации грамматики финно-угорских языков

предложено использовать метод на основе модификации формальных грамматик.

Библиографический список

[Килеев и др., 2011] Килеев, В.В. Лингвистические особенности архитектуры компьютерной системы верификации орфографии финно-угорских языков / В.В. Килеев, И.Г. Сидоркина // Вестник Волжского университета имени В.Н. Татищева. Научно-теоретический журнал. – Серия «Информатика». – Вып. 18. – Тольятти: Волжский университет им. В.Н. Татищева, 2011г. – С. 115-119.

[Bender, et al., 2004] Bender E. [и др.] Arboretum: Using a precision grammar for grammar checking in CALL [Конференция] // In Proceedings of the InSTIL/ICALL Symposium. - 2004.

[Berlinsky-Schine, 2004] Berlinsky-Schine A. Context-based detection of 'real word' typographical errors using markov models [Report] : Technical report. - Ithaca, NY : Cornell University, 2004.

[Budanitsky, et al., 2006] Budanitsky A. and Hirst G. Evaluating wordnet-based measures of lexical semantic relatedness [Journal] // Computational Linguistics. - 2006. - 32(1). - pp. 13-47.

[Fossati, et al., 2007] Fossati D. and Di Eugenio B. A mixed trigrams approach for context sensitive spell checking [Conference] // CICLing-2007, Eighth International Conference on Intelligent Text Processing and Computational Linguistics. - Mexico City, Mexico : [s.n.], 2007.

[Fossati, et al., 2008] Fossati D. and Di Eugenio B. I saw TREE trees in the park: How to correct real word spelling mistakes. [Conference] // LREC 2008, 6th International Conference on Language Resources and Evaluation. - Marrakech, Morocco : [s.n.], 2008.

[Gale, et al., 1993] Gale W. A., Church K. W. and Yarowsky D. A method for disambiguating word [Journal] // Computers and the Humanities. - 1993. - 26. - pp. 415-439.

[Garside, et al., 1987] Garside R., Leech G. and Sampson G. The Computational Analysis of English: a corpus based approach [Book]. - [s.l.] : Longman, 1987.

[Golding, 1995] Golding A. A bayesian hybrid method for context-sensitive spelling correction [Journal] // In The Third Workshop on Very Large Corpora. - 1995. - pp. 39–53.

[Golding, et al., 1996] Golding A. and Schabes Y. Combining trigram-based and feature-based methods for contextsensitive spelling correction [Conference] // 34th Annual Meeting of the Association for Computational Linguistics. - 1996.

[Heidorn, et al., 1986] Heidorn G. E. [et al.] The EPISTLE text-critiquing system [Journal] // IBM Systems Journal. - 1986. - 21(3). - pp. 305–326.

[Hirst, et al., 2005] Hirst G. and Budanitsky A. Correcting real-word spelling errors by restoring lexical cohesion [Journal] // Natural Language Engineering. - 2005. - 11. - pp. 87-111.

[Hirst, 2008] Hirst G. An Evaluation of the Contextual Spelling Checker [Доклад]. - Toronto, Canada : Department of Computer Science, 2008.

[Jones, et al., 1997] Jones M. P. and Martin J. H. Contextual spelling correction using latent semantic analysis [Conference] // Fifth Conference on Applied Natural Language Processing. - 1997.

[Marshall, 1983] Marshall I. Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus // Computers and the Humanities. - 1983. - 17. - pp. 139-150.

[Mays, et al., 1991] Mays E., Damerau F. J. and Mercer R. L. Context based spelling correction [Journal] // Context based spelling correction. Information Processing and Management. - 1991. - 27(5). - pp. 517–522.

[Naber, 2003] Naber D. A Rule-Based Style and Grammar Checker. Diploma thesis, University of Bielefeld.

[Wilcox-O'Hearn, et al.] Wilcox-O'Hearn A., Hirst G. and Budanitsky A. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. [Conference] // CICLing-2008, 9th.

[Yarowsky, 1994] Yarowsky D. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French // Proceedings of the 32nd Annual Meeting of the Association for computational Linguistics. - 1994. - pp. 88–95.

NATURAL FINNO-UGRIC LANGUAGE GRAMMAR VERIFICATION METHODS AT THE LEVEL OF SEMANTIC REPRESENTATION

Kileev V.V.*, Sidorkina I.G.*

**Volga State University of Technology, Yoshkar-Ola, Republic of Mari El, Russia*

slavakileev@yandex.ru

igs592000@mail.ru

In this paper review of natural language grammar verification methods is considered in the context of semantic representation to implement it in the system. In the work of grammar verification system it is required to determine semantics of the verified word to check it for correspondence to the context. The difference between grammar verification methods is lied in the ways of determining the verified word's semantics. The ways of determining verified word's semantics are examined in addition to the short description of each method

INTRODUCTION

In the field of natural language grammar verification one of the main targeted task is to determine meaning (semantics) of the verified text. The difference between grammar and spell checking is that in grammar checking the correspondence of certain word to the context is determined. Thereby grammar verification determines semantics of the word and then checks whether it corresponds to semantics of the whole sentence.

MAIN PART

The fragment of general classification of grammar verification methods is given on the fig. 1. There are three main types of grammar verification methods: Symbol methods, Statistical methods and methods that use formal grammar.

In symbol methods templates for each mistake type is manually created. The various statistical methods require large text corpora to work correctly.

Methods that use formal grammar may use one grammar to store all grammatical rules and may use several grammars with different level of detail. In the second case the logical inference is done on all grammars simultaneously. The error is detected when the results of the inference on different grammars are not equal.

CONCLUSION

Though statistical methods are widely used nowadays they require large text corpora. Because of the fact that for Eastern Finno-Ugric languages there are no such text corpora it is proposed to use methods based on modified formal grammars in creation of Finno-Ugric languages grammar verification system.