КРАТКИЕ СООБЩЕНИЯ

СИСТЕМНЫЙ АНАЛИЗ И ОБУЧАЮЩИЕ СИСТЕМЫ

УДК 004.853

Я. А. Жук
Белорусский государственный технологический университет

ОЦЕНКА ЭФФЕКТИВНОСТИ РАБОТЫ ГЕНЕРАТОРА СЕМАНТИЧЕСКОЙ СЕТИ ДИАЛОГОВОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Статья посвящена методике расчета эффективности работы генератора семантической сети диалоговой информационной системы. Данная методика предназначена для отладки алгоритма выявления знаний в HTML-коде электронных учебников и наполнения базы знаний диалоговой системы. Разработанная методика основывается на экспертных оценках качества обработки естественного языка. При расчете эффективности учитывается присутствие в электронных учебниках вспомогательных и сложных предложений. Выявлены и классифицированы некоторые ошибки функционирования разработанного генератора семантической сети информационной системы. Установлено, что данные ошибки связаны с применением в содержании электронных учебников сложных синтаксических конструкций (списков, отсылок) и размещением знаков препинания без учета HTML-разметки. Предложены направления решения описанных проблем. В качестве основного пути решения предложено внесение в алгоритм работы генератора семантической сети диалоговой информационной системы дополнительных функций для обработки списков и отсылок, а также корректного выбора сказуемого в предложениях. В качестве дополнительного способа решения проблем генерации семантической сети предложена подготовка правил предварительной корректуры HTML-кода компьютерных обучающих систем. Соблюдение данных правил обеспечит корректное выполнение морфологического и синтаксического анализа содержания электронных учебников.

Ключевые слова: компьютерная обучающая система, обработка естественного языка, семантическая сеть, диалоговая система, оценка качества.

Ya. A. Zhuk
Belarusian State Technological University

EFFICENCY MEASUREMENT OF THE DIALOG INFORMATION SYSTEM SEMANTIC NETWORK GENERATOR

The article describes the method of the efficiency calculating of the dialog information system semantic network generator in order to debug the algorithms of knowledge detection. This technique is designed to debug the algorithm of knowledge mining from the HTML-code of electronic textbooks and filling the knowledge base of the dialogue system. The developed method is based on expert assessments of the natural language processing quality. When calculating the effectiveness, the presence of auxiliary and complex sentences in electronic textbooks is taken into account. Some errors in functioning of the developed information system semantic network generator are revealed and classified. It is established that these errors are associated with the use of complex syntactic structures (lists, references) in the content of electronic textbooks and the placement of punctuation marks without taking HTML-markup into account. The directions of the solution of the described problems are offered. As the main way of solution it is proposed to add additional functions into the algorithm of the dialog information system semantic network generator for processing lists and references, as well as the correct choice of the predicate in the sentences. As an additional way to solve the problems of semantic network generation, the preparation of rules for preliminary proofreading of HTML-code of computer

training systems is proposed. Compliance with these rules will ensure the correct implementation of morphological and syntactic analysis of the electronic textbooks content.

Key words: e-learning system, natural language processing, semantic network, dialog system, quality measurement.

Введение. Существующие электронные учебники состоят из различных элементов. Текст в таких системах содержит предложения различной структуры, уточнения в скобках, отсылки к предыдущим предложениям, различные виды оборотов в сложноподчиненных предложениях [1]. Кроме обычного текста встречаются списки, изображения, таблицы, звукозаписи, видеоролики, интерактивный контент. При автоматическом извлечении знаний из такого смешанного содержания требуется выполнять детальный анализ каждого фрагмента для применения соответствующего алгоритма обработки. Кроме того, наличие в тексте различных элементов следует учитывать при оценке эффективности работы генератора семантической сети для выявления ситуаций, требующих дополнительной отладки [2]. Важно отметить, что в литературе встречаются методики оценки эффективности программ обработки естественного языка, но они предназначены для оценки морфологического и синтаксического анализа, а не семантического [3]. Далее предлагается методика оценки эффективности работы генератора семантической сети диалоговой информационной системы.

Основная часть. Оценку эффективности работы генератора семантической сети предлагается производить путем сравнения содержания электронного учебника с результатом анализа данного содержания генератором семантической сети. Как и в других методах оценки эффективности программ обработки естественного языка, сравнение выполняется экспертом [3]. Важно подчеркнуть, что эксперт действует только на этапе отладки генератора семантической сети, после отладки программный модуль будет работать автоматически.

В качестве информационных систем, исследуемых на эффективность разработанного генератора семантической сети, выбраны достаточно разнородные по содержанию электронные учебники по дисциплинам «Электрохимия» и «Микробиология», представленные в HTML-формате.

Эффективность работы генератора семантической сети Э предлагается рассчитывать по формуле

$$Э = \frac{C_К + B_К}{C_О + B_О},$$

где $C_К$ – количество корректно выявленных семантических связей; $B_К$ – количество корректно выявленных вспомогательных предложений; $C_О$ – общее количество семантических связей, которое возможно выявить в тексте; $B_О$ – общее количество выявленных экспертом вспомогательных предложений.

Такая методика принимает во внимание следующие важные моменты, связанные с применением семантической сети в качестве базы знаний для функционирования диалогового модуля [4]:

– не все предложения в электронных учебниках сами по себе несут знания, которые можно использовать в качестве ответа на вопрос, поэтому выявление таких вспомогательных предложений при отсутствии обнаруженных семантических связей засчитывается как положительный результат, позволяющий избежать засорения семантической сети некорректной информацией и по ценности принимаемый равным одной семантической связи;

– хотя выявленная семантическая связь может иметь крайне низкую вероятность быть выбранной в качестве самостоятельного ответа на вопрос из-за низкой вероятности ввода учениками данного вопроса, следует засчитывать ее как положительный результат работы, так как данная связь может быть использована вместе с другими в расширенном ответе.

При расчете показателей, требуемых для оценки эффективности по данной методике, предлагается использовать циклический подход. В начале расчета все показатели следует считать равными нулю. Цикл выполняется над предложениями исходного содержания информационной системы. На каждой итерации цикла текущее предложение сравнивается с набором семантических связей, выявленных генератором семантической сети в данном предложении. Эксперт продумывает, какие семантические связи можно извлечь из предложения и увеличивает $C_О$ на количество данных связей. При наличии схожих семантических связей в результатах анализа $C_К$ увеличивается на количество таких связей. Если же эксперт считает, что предложение является вспомогательным и семантические связи из него извлечь невозможно, то $B_О$ увеличивается на единицу. Если при этом генератор семантической сети также не выявил семантических связей, то $B_К$ увеличивается на единицу.

Помимо подсчета семантических связей с целью отладки генератора семантической сети был проведен подсчет и классификация связей, выявленных экспертом, но не выявленных либо выявленных некорректно генератором семантической сети. Было выявлено 5 типов таких ошибок.

1. Некорректное восполнение недостающих членов предложения генератором семантической сети и некорректное определение сказуемого в сложноподчиненных предложениях. Например, при объяснении происхождения слова «*комплемент дополняет (от лат. complementum – дополнение)*» уточнение в скобках было разобрано как определение и образовало следующую семантическую связь: «*('от латинского complementum', 'дополнение', 'что обознача[Г1]ся [К] [А]')*». Также и в случае сложноподчиненных предложений с несколькими сказуемыми выбор не всегда осуществляется правильно, например, в предложении «*МАС проникает через плазматическую мембрану бактериальной клетки, создавая в ней бреши, в результате чего клетка подвергается осмотическому лизису*» была выявлена следующая семантическая связь «*('МАС проникает через плазматическую мембрану бактериальной клетки, создавая в ней бреши, в результате чего клетка', 'осмотическому лизису', '[В] подверга[Г1]ся [А]', NULL)*», т. е. в качестве сказуемого был выбран глагол «*подвергается*», а не «*проникает*».

2. Неразрешенные отсылки к предыдущему предложению, иногда к более ранним предложениям, выраженные местоимениями и другими частями речи, например «*Понятие обратимо работающего ГЭ не следует путать с понятием обратимого ГЭ. Второе понятие, как мы только что видели, определяется химической природой электродов, из которых состоит данный элемент. Первое же понятие связано с условиями работы обратимого элемента*».

3. Некорректная обработка списков, которые могут представлять собой набор однотипных связей или связей различных типов, например: «* Для создания вакцин методами генетической инженерии используют следующие подходы: – осуществляют химический синтез пептидов, обладающих антигенными свойствами в составе того или иного патогена, выяснив предварительно аминокислотную последовательность такого белка; – встраивают гены, кодирующие структуру <a onClick = "Book.loadPageWithHehghtleight ('thesaurus', 'antigen_determ'); ">антигенных детерминант , в векторные молекулы, которые переносят в клетки непатогенных микроорганизмов или в растения и обеспечивают синтез заданных продуктов, используя их затем в качестве вакцин; *».

4. Некорректное выявление границ предложений генератором семантической сети, в частности из-за попадания точек внутрь тегов, рассматриваемых как единое целое, например «*Одним из очень важных защитных барьеров организмов животных и человека от патогенов является нормальная микробиота, состав и роль которой рассмотрены в <a onClick = "Book.loadChapterBack(99);"> § 18.3. *» – точка после номера параграфа оказалась внутри тега <a>... и поэтому не была распознана как конец предложения.

5. Предложения построены не по шаблону «подлежащее – сказуемое – вспомогательные члены предложения». Примером таких предложений являются причинно-следственные связи, например «*Когда иммунный ответ на внедрение какого-то антигена состоялся, то включилась система «памяти», которая способна обеспечить быстрый специфический вторичный ответ*». Еще одним примером таких предложений являются перечисления существующих разновидностей объектов. Например «*Различают активный и пассивный иммунитет*».

В ходе предварительного анализа отдельных параграфов электронных учебников по электрохимии и микробиологии была обнаружена достаточно низкая эффективность работы генератора семантической сети: 35% при анализе параграфа электронного учебника по электрохимии и 42% при анализе параграфа электронного учебника по микробиологии. Было обнаружено, что наибольшее количество связей теряется из-за ошибок 2-, 3- и 4-го типов. Путем внесения в алгоритм генератора семантической сети дополнительных условий и действий, направленных на обработку перечислений, отсылок к предыдущему предложению и устранение объединения предложений, удалось значительно повысить показатель эффективности. Результаты обработки тех же текстов с помощью улучшенного генератора семантической сети представлены в таблице.

Оценка результатов анализа кода учебников

Показатель	Учебник «Электрохимия», § 3.2	Учебник «Микробиология», § 18.4
$C_К$	29	68
$B_К$	5	4
$C_О$	56	107
$B_О$	5	4
Э	0,55	0,64
Ошибки типа 1	9	17
Ошибки типа 2	11	9
Ошибки типа 3	3	1
Ошибки типа 4	3	7
Ошибки типа 5	1	5

Как видно из таблицы, эффективность работы генератора семантической сети превысила половину и составила 55% при анализе параграфа электронного учебника по электрохимии и 64% при анализе параграфа электронного учебника по микробиологии.

Значительная часть оставшихся ошибок связана с тем, что в анализируемых текстах присутствуют отсылки не только к предыдущему предложению, но и к более ранним. Также были замечены отсылки, которые невозможно корректно разрешить на основании морфологических признаков слов и их порядка в предложениях. Для охвата различных видов отсылок требуется система отслеживания контекста и предварительная подготовка семантической сети базовых понятий предметной области, используемых в тексте.

Стоит отметить, что еще одной значительной группой ошибок является некорректное дополнение и объединение блоков с целью получения полных простых и сложноподчиненных предложений. Для устранения ошибок данного вида требуется разработка дополнительных условий и действий для корректной обработки широкого спектра специфических ситуаций.

Также следует подчеркнуть, что эффективность работы генератора семантической сети в значительной мере зависит от анализируемого текста. Это говорит о том, что некоторые из ошибок могут быть устранены форматированием HTML-кода электронных учебников в соответствии с дополнительными правилами, такими как вынесение точек, обозначающих конец предложения, за пределы тегов.

Заключение. Описанный подход к расчету эффективности извлечения знаний позволил выявить необходимость доработки генератора семантической сети для корректного анализа отсылок к предыдущим предложениям, списков, причинно-следственных связей и обобщенно-личных предложений. В соответствии с подсчетами корректный анализ отсылок наиболее важен, поскольку они широко распространены в электронных учебниках. Также значительно повысить эффективность может корректный анализ списков и правильное выявление границ предложений.

Литература

1. Стилистический энциклопедический словарь русского языка / под ред. М. Н. Кожиной. М.: Флинта: Наука, 2011. 696 с.

2. Гурин Н. И., Жук Я. А. Генератор семантической сети информационной системы в таблицу реляционной базы данных // Труды БГТУ. 2015. № 6: Физ.-мат. науки и информатика. С. 181–185.

3. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка / С. Ю. Толдова [и др.] // Компьютерная лингвистика и интеллектуальные технологии. 2012. Вып. 2, № 11. С. 77–90.

4. Жук Я. А., Гурин Н. И. Реализация диалога с компьютерной обучающей системой на языке JavaScript с помощью веб-сервисов // Труды БГТУ. Сер. 3, Физ.-мат. науки и информатика. 2018. № 2. С. 107–112.

References

1. *Stilisticheskiy entsiklopedicheskiy slovar'* [Stylistic encyclopedic dictionary]. Moscow, Flinta Publ., Nauka Publ., 2011. 696 p.

2. Gurin N. I., Zhuk Ya. A. The information system semantic network generator to a relational database table generator. *Trudy BGTU* [Proceedings of BSTU], 2015, no. 6: Physics and Mathematics. Informatics, pp. 181–185 (In Russian).

3. Toldova S. Yu., Sokolova E. G., Astaf'eva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O. N. Evaluation of methods for automatic text analysis 2011–2012: Syntactic parsers of the russian language. *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii* [Computational linguistics and intellectual technologies], 2012, issue 2, no. 11, pp. 77–90 (In Russian).

4. Zhuk Ya. A., Gurin N. I. Implementation of dialogue with the computer training system in JavaScript using web services. *Trudy BGTU* [Proceedings of BSTU], series 3, Physics and Mathematics. Informatics, 2018, no. 2, pp. 107–112 (In Russian).

Информация об авторе

Жук Ярослав Александрович – аспирант кафедры информационных систем и технологий. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: zhuk@belstu.by

Information about the author

Zhuk Yaroslav Aleksandrovich – PhD student, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: zhuk@belstu.by

Поступила 15.02.2019