Инженерный вестник Дона, №5 (2022)
ivdon.ru/ru/magazine/archive/n5y2022/7673

Методы определения целевого предложения для автоматизированной генерации тестовых вопросов

М.А. Маслова

Волгоградский государственный технический университет

Аннотация: Автоматизация генерации тестовых вопросов состоит из четырех основных этапов. Первым этапом является определение целевого предложения, из которого можно будет сгенерировать тестовый вопрос. В данной статье исследуются существующие методы по определению целевого предложения. Рассматриваются как иностранные, так и русскоязычные источники. В работе проанализированы такие методы, как автоматическое реферирование тестов, сходство синтаксического анализа, метод, основанный на правилах, метод, основанный на ранжировании, по ключевым словам, алгоритм обобщения документов, основанный на частотности слов.
Ключевые слова: автоматическая генерация тестовых вопросов, автоматическая обработка текста, обработка естественного языка, автоматическое реферирование.

Введение

В последнее время преподаватели часто используют тестирование для проверки теоретических или практических знаний. Но разработка тестовых заданий требует затрат большого количества времени и большого количества труда преподавателя. Поэтому многие преподаватели или научные исследователи интересуются автоматической генерацией тестовых заданий на основе текстовой информации.

Автоматическую генерацию вопросов можно поделить на четыре шага: определение целевых предложений, из которых можно сгенерировать тестовое задание, определение слова или фразы, которое будет целевым ответом, преобразование предложения в вопрос, генерация неверных вариантов ответа. Определение целевого предложения является важным этапом в автоматической генерации тестовых вопросов, так как от этого этапа зависит качество сгенерированных вопросов. Для того, чтобы сгенерировать корректные вопросы, необходимо выбирать предложения, имеющие информационную значимость.

Целью данной статьи является исследование существующих методов определения целевого предложения, из которого можно сгенерировать тестовое задание, и выбрать оптимальный метод для использования в системе генерации вопросов.

Основная часть

Идея автоматической генерации тестовых вопросов [1] рассматривается во многих научных работах. В литературе описываются различные методы для определения целевого предложения.

Так, Куртасов и Швецов [2] рассматривают для выделения целевых предложений автоматическое реферирование текста. Авторы пишут, что у данного метода в области автоматической обработки текста есть два основных подхода: выделение важных предложений и формирование аннотации (аннотирование) [3]. Было решено использовать первый подход, выделение важных предложений. Данный выбор обоснован тем, что реализация второго подхода более трудоемкая, а также подчеркнули, что для решаемой задачи главной целью является удалить из исходного текста неинформативные предложения, для чего, собственно, больше подходит первый подход. Но авторы не используют автореферирование в исследуемом прототипе, однако для генерации вопросов применяются некоторые правила, которые позволяют отбирать предложения определенной структуры, например, определения терминов.

M. Majumder и S. K. Saha [4] используют сходство синтаксического анализа. Авторы рассматривают вопросы как простые предложения, но они столкнулись со следующей проблемой: многие предложения в Википедии и новостных статьях длинные, сложносочиненные и сложноподчинённые. Более того, в ряде этих предложений возникают проблемы с совместными ссылками. Предлагаемая методика основана на сходстве структуры синтаксического анализа; следовательно, структура предложений играет

важную роль в задаче. Чтобы получить лучшее структурное сходство, они сначала применяют несколько этапов предварительной обработки.

Чтобы оценить производительность системы, они взяли несколько страниц Википедии и новостных статей в качестве входных данных, на которых они запустили систему. Способность полученных предложений к формированию вопросов проверяется набором оценщиков-людей. Оценщики подсчитывают количество предложений, которые потенциально могут стать целевым предложением для генерации вопроса (‘правильный поиск’). Среднее значение процента правильного поиска считается точностью системы.

Точность системы, основанной на сходстве синтаксического дерева.

Для вычисления точности системы авторы рассматривают шесть страниц Википедии. В качестве входных данных берутся только текстовые части этих страниц, которые содержат в общей сложности ~ 795 предложений. Из этого входного текста было выбрано ~ 508 предложений после фильтрации на основе тематических слов. Затем авторы применяют алгоритм сопоставления дерева синтаксического анализа, который в конечном итоге рассматривает 112 предложений. Эти предложения рассматриваются пятью экспертами-оценщиками. Они рассматривают 105, 104, 103, 106 и 104 предложения соответственно, как правильный поиск. Таким образом, точность системы составляет 93,21%.

B. Das и M. Majumder [5] предложили метод для извлечения информативных предложений основанный на правилах.Анализ тегов части речи в предложении является основой предлагаемых ими правил. Чтобы определить информативные предложения, авторы [6] сначала собрали простые предложения из анализа зависимостей входного корпуса. Далее они проанализировали простые предложения и рассмотрели те, которые не

превышают 20 слов и не содержат тегов RB/RBR/RBStag (наречие) и с наименьшими двумя объединенными тегами NNP/NNPS (имя собственное). Предложения, обладающие вышеупомянутыми свойствами, дополнительно уточняются на основе правил, основанных на тегировании частей речи.

Чтобы проверить точность предлагаемой системы, авторы извлекли данные из восьми страниц Википедии. Эти страницы содержат в общей сложности около 2275 предложений; из них 614 являются простыми предложениями. Следовательно, 614 предложений передаются системе, которая идентифицирует 131 предложение как информативное. Поскольку не существует стандарта для вычисления точности такой системы, они взяли суждения пяти экспертов-лингвистов о правильности извлеченных предложений и рассмотрели точность как среднее значение их суждений. Они рассмотрели 120, 118, 122, 124 и 121 предложения соответственно, как приемлемые информативные простые предложения. Следовательно, точность предложенной системы составляет 92,367%.

L. Becker, S. Basu and L. Vanderwende [7] используют их собственную реализацию SumBasic [8], алгоритма обобщения документов, основанного на предположении, что предложения, содержащие наиболее часто встречающиеся слова в статье, являются наиболее важными. Таким образом, авторы используют суммарную оценку для каждого предложения, чтобы упорядочить их в качестве кандидатов для построения вопроса.

SumBasic сначала вычисляет распределение вероятности по словам, появляющимся на входе. При вычислении распределения вероятностей учитываются только глаголы, существительные, прилагательные и числа. Далее присваивается вес важности каждому предложению во входных данных в зависимости от важности слов его содержания. И наконец, система выбирает предложение с наилучшей оценкой в соответствии с функцией

оценки с предыдущего шага. Если требуемая длина сводки не достигнута, система возвращается присваиванию веса важности.

B. Das, M. Majumder, S. Phadikar и A. A. Sekh [9] определяют целевые предложения с помощью ключевых слов. Они отделили все существующие простые предложения от других предложений [10], используя идентификацию одного независимого предложения в предложении, используя метод, описанный в [5]. Далее, вес полученного предложения рассчитывается путем объединения весов отдельных ключевых слов, принадлежащих предложению. Авторы присваивают более высокий вес ключевому слову, в котором больше слов. Вес ключевого слова в предложении определяется количеством отдельных слов, присутствующих в ключевом слове. Предложения с наивысшим рейтингом выбираются в качестве информативных предложений для создания вопросов.

Выводы

Согласно рассмотренным работам, большинство использует суммаризацию текста. Аннотирование позволяет удалить из текста самые неинформативные предложения, но также многие применяют ранжирование предложений в тексте по ключевым словам. Ранжирование помогает выявить предложения содержащие больше значимой и важной информации. Поэтому стоит обратить внимание на эти два способа выявления информативных предложений для того, чтобы применить их к определению целевых предложений, которые будут использоваться в автоматической генерации тестовых вопросов.

На основе проведённого анализа, планируется использовать автореферирование текста и ранжирование предложений, по ключевым словам, для определения целевого предложения в автоматической генерации тестовых вопросов на основе текстовой информации.

Литература

1. Алсынбаева Л.Г. Система автоматизированной генерации тестовых заданий // Программные продукты и системы. 2009. №4 (2). URL: cyberleninka.ru/article/n/sistema-avtomatizirovannoy-generatsii-testovyh-zadaniy/viewer.

2. Куртасов А.М., Швецов А.Н. Метод автоматизированной генерации заданий для тестов контроля знаний из текстов учебных пособий // Современные информационные технологии и ИТ-образование. 2013. №9. URL: elibrary.ru/item.asp?id=23020528.

3. Nenkova A., McKeown K. Automatic Summarization // Foundations and Trends in Information Retrieval. 2011. №5. Pp. 103-233.

4. Majumder M., Saha S.K. A System for Generating Multiple Choice Questions: With a Novel Approach for Sentence Selection // Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications. 2015. Pp. 64-72.

5. Das B., Majumder M. Factual open cloze question generation for assessment of learner's knowledge // International Journal of Educational Technology in Higher Education. 2017. №1(14). URL: educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-017-0060-3

6. Santorini B. Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision) // ScholarlyCommons. 1990. 32 p.

7. Becker L., Basu S., Vanderwende L. Mind the Gap: Learning to Choose Gaps for Question Generation // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012. Pp. 742-751.

8. Nenkova A., Vanderwende L., McKeown K. A compositional context sensitive multi-document summarizer: exploring the factors that influence

summarization // Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006. URL: semanticscholar.org/paper/A-compositional-context-sensitive-multi-document-Nenkova-Vanderwende/6d9f0bcb44ac9a266e5de7eeada0f01b5f957d84

9. Das B., Majumder M., Phadikar S., Sekh A.A. Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment // Multimedia Tools and Applications. 2021. №21(80). Pp. 31907-31925

10. Das B., Majumder M., Phadikar S. A novel system for generating simple sentences from complex and compound sentences // Int J Modern Educ Comput Sci. 2018. №1(10). Pp. 57-64

References

1. Alsynbaeva L.G. Programmnye produkty i sistemy. 2009. №4(2). URL: cyberleninka.ru/article/n/sistema-avtomatizirovannoy-generatsii-testovyh-zadaniy/viewer.

2. Kurtasov A.M., SHvecov A.N. Sovremennye informacionnye tekhnologii i it-obrazovanie. 2013. №9. URL: elibrary.ru/item.asp?id=23020528.

3. Nenkova A., McKeown K. Foundations and Trends in Information Retrieval. 2011. №5. Pp. 103-233.

4. Majumder M., Saha S.K. Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications. 2015. C. 64-72.

5. Das B., Majumder M. International Journal of Educational Technology in Higher Education. 2017. №1 (14). URL: educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-017-0060-3

6. Santorini B. Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision). ScholarlyCommons. 1990. 32 Pp.

7. Becker L., Basu S., Vanderwende L. Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012. Pp. 742-751.

8. Nenkova A., Vanderwende L., McKeown K. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006. URL: semanticscholar.org/paper/A-compositional-context-sensitive-multi-document-Nenkova-Vanderwende/6d9f0bcb44ac9a266e5de7eeada0f01b5f957d84

9. Das B., Majumder M., Phadikar S., Sekh A.A. Multimedia Tools and Applications. 2021. №21 (80). Pp. 31907-31925

10. Das B., Majumder M., Phadikar S. Int J Modern Educ Comput Sci. 2018. №1 (10). Pp. 57-64.