УДК 025.4.036+004.912

ВЗВЕШЕННАЯ ПОГРЕШНОСТЬ – НОВАЯ МЕТРИКА ДЛЯ ОЦЕНКИ КАЧЕСТВА ВАЛИДАЦИИ ОТВЕТОВ В ЗАДАЧЕ ВОПРОСНО-ОТВЕТНОГО ПОИСКА

А.А. Соловьев

МГТУ им. Н.Э. Баумана, Москва
e-mail: a-soloviev@mail.ru

Рассмотрена подзадача валидации ответов в задаче вопросно-ответного поиска. Традиционными метриками качества на семинарах TAC-RTE и CLEF-AVE являются аккуратность (accuracy) и F-мера. По результатам участия в семинаре РОМИП-2010 отмечено, что число ложных ответов-гипотез, которые должен отклонить модуль валидации ответов, часто значительно превышает число верных ответов. Предложена новая метрика — взвешенная погрешность, которая чаще штрафует систему за ошибки первого рода (пользователю показан неверный ответ — falsepositive), чем за ошибки второго рода (правильный ответ отвергнут и пользователю не показан — falsenegative). В отличие от F-меры она также поощряет систему за верно отфильтрованный ответ (truenegative).

Ключевые слова: информационный поиск, вопросно-ответный поиск, вопросно-ответные системы, проверка ответов, валидация ответов, компьютерная лингвистика, обработка естественного языка.

WEIGHTED ERROR – NEW METRICS FOR ESTIMATING QUALITY OF ANSWER VALIDATION IN THE PROBLEM OF QUESTION-ANSWERING RETRIEVAL

A.A. Solovyev

Bauman Moscow State Technical University, Moscow
e-mail: a-soloviev@mail.ru

The answer validation subproblem is considered in a problem of question answering retrieval. Traditional quality metrics at the TAC-RTE and CLEF-AVE seminars are accuracy and F-measure. From results of participation in ROMIP-2010 seminar, it is noted that a number of false answer-hypotheses that must be declined by the answer validation module frequently exceeds substantially the true answer number. A novel metrics—weighted error is proposed which penalizes the system for the first-kind errors (false positive errors, when an incorrect answer is shown to the user) more frequently than for the second-kind errors (false negative errors, when a correct answer is rejected and not shown to the user). Unlike the F-measure, it also rewards the system for the properly rejected (true negative) answer.

Keywords: information retrieval, question answering, answer validation, question-answering systems, checking answer, answer validation, computational linguistics, natural language processing.

Программные системы вопросно-ответного поиска, или просто вопросно-ответные системы (*англ.* QuestionAnsweringSystems) – это

вид информационно-поисковых систем, способных обрабатывать введенный пользователем вопрос на естественном языке и выдавать осмысленный ответ. В отличие от задачи классического поиска по ключевым словам, в которой результатом является перечень документов, содержащих ответ на вопрос, в задаче вопросно-ответного поиска — это краткий и лаконичный ответ, сформированный системой в результате анализа разнообразных источников данных. Примером такого источника может служить некоторая коллекция полнотекстовых документов (множество страниц глобальной сети Интернет), а ответ составляется из фрагмента наиболее релевантного документа коллекции.

Обзор существующих методов валидации ответов, описание разрабатываемого метода параллельного обхода графов и формулировка задачи экспериментального исследования этих методов были приведены в работе [1]. В настоящей статье рассмотрены существующие подходы к экспериментальной оценке качества вопросно-ответных систем, в частности модуля валидации ответов. Обоснован выбор новой метрики для выполнения экспериментов, заявленных в работе [1].

Оценка вопросно-ответной системы в целом. Для оценки вопросно-ответной системы в целом применяются следующие метрики:

- Mean reciprocal rank [2];
- Confidence weighted score [3];
- Аккуратность [4];
- NIL-точность и NIL-полнота [3];
- c@1 [5].

Оценить валидацию ответа при таком подходе можно, сравнивая прогоны системы в разных конфигурациях:

- с отключенным модулем валидации;
- с тривиальной реализацией модуля (например, на модели мешка слов);
- с вырожденной реализацией (отклонять все ответы);
- с реализацией методов, предложенных другими авторами;
- с предлагаемой реализацией, но с разными параметрами.

Сравнивая результаты этих прогонов можно оценить вклад предлагаемой реализации модуля валидации ответов в качество вопросно-ответной системы в целом.

Важным требованием к методу оценки системы в целом является возможность учета варианта “нет ответа”. Чтобы вычислить такие метрики, как *NIL-точность* и *NIL-полнота*, необходимо знать, есть ли вообще в данной коллекции документов ответ на каждый тестовый вопрос. Обычно такая информация добывается методом общего котла: если хоть одна из тестируемых систем дала правильный ответ

на вопрос (т.е. отмеченный асессорами как правильный), то ответ на вопрос существует.

Таким образом, если оцениваемая система дает неверный ответ на вопрос, для которого не существует ответа, значит у нее *низкая NIL-полнота*. Если система не дает ответ на вопрос, для которого какая-то другая система успешно нашла ответ, значит у нее *низкая NIL-точность*.

Процедура оценки вопросно-ответной системы очень трудоемка, так как требует работы нескольких асессоров, оценивающих результаты множества прогонов. Обычно такую оценку проводят в рамках ежегодных кампаний TREC, CLEF, TAC, РОМИП.

В работе [6] опубликованы результаты участия автора в семинаре по оценке методов информационного поиска РОМИП. Организаторы семинара отметили низкую эффективность кампании (дорожки вопросно-ответного поиска в 2010 г.) — значительные усилия асессоров были потрачены, чтобы констатировать тот факт, что тестовые вопросы слабо соответствовали предложенной участникам коллекции документов. Так, только для 60 заданий из 246 асессоры предполагали, что документ с ответом существует в коллекции.

К сожалению, метрики вопросно-ответной дорожки, предоставленные организаторами, не позволяли адекватно сравнить прогоны, так как никак не поощряли вариант “нет ответа”, хотя это должен быть самый распространенный правильный ответ. Были представлены следующие метрики:

- на сколько запросов был подан хоть один вариант ответа;
- число запросов у которых есть хотя бы один ответ с оценкой *good*;
- число запросов, у которых есть хотя бы один ответ с оценками *good*, *long* или *partial*;
- число запросов, у которых есть хотя бы один длинный ответ (фрагмент) с оценкой *good*;
- число запросов, у которых есть хотя бы один длинный ответ (фрагмент) с оценками *good* или *partial*.

Чтобы исправить этот недочет, были предложены две метрики на основе категорий вопросов, представленных в табл. 1:

ошибка E — отношение числа неправильно принятых решений к общему числу решений,

$$E = \frac{b + c + d}{a + b + c + d + e};$$

полнота R — отношение числа вопросов с правильными ответами к общему числу вопросов, имеющих ответ в коллекции,

$$R = \frac{a}{a + b + d}.$$

Таблица 1

Предложенные категории ответов системы для заданий РОМИП 2010 (не являются официальными метриками РОМИП)

Прогон	Эталон	
	Ответ на вопрос есть в коллекции	Правильного ответа на вопрос в коллекции нет
Система дала хотя бы один правильный ответ на вопрос	a	0
Система дала один или несколько ответов на вопрос, но все неправильные	b	c
Система не дала ни одного ответа на вопрос	d	e

Результаты эксперимента РОМИП показали, что применение предложенного метода валидации ответов позволило снизить уровень ошибок E с 59 до 26 % при снижении полноты R с 8 до 5 %.

После участия в РОМИП было принято решение построить тестовую коллекцию вопросов и ответов на основе заданий РОМИП, но используя другой источник текстов — поисковую выдачу Яндекса. Наличие такой коллекции с положительными и отрицательными примерами ответов позволяет выполнять воспроизводимые эксперименты для задачи валидации ответов, но не для оценки вопросно-ответной системы в целом.

Валидация ответов как задача бинарной классификации. Предлагаемый способ оценки валидации ответов основан на традиционном подходе к оценке в задаче классификации. Рассмотрим задачу валидации как задачу бинарной классификации: тройку ⟨вопрос, ответ, сниппет⟩ требуется отнести к одному из двух классов — верный ответ (правильность ответа на вопрос следует из предоставленного сниппета) или неверный.

В табл. 2 приведены четыре возможных исхода решения задачи классификации.

Таблица 2

Категории результата бинарной классификации ответов

Наблюдаемый результат	Ожидаемый результат	
	Верный ответ	Неверный ответ
Верный ответ	tp (true-positive)	fp (false-positive, ошибка первого рода)
Неверный ответ	fn (false-negative, ошибка второго рода)	tn (true-negative)

На основе этой таблицы определяются традиционные метрики качества классификации:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn};$$

$$Error = \frac{fp + fn}{tp + tn + fp + fn} = 1 - Accuracy;$$

$$Precision = \frac{tp}{tp + fp}; \quad Recall = \frac{tp}{tp + fn};$$

$$F_\beta = \frac{(1 + \beta^2) Precision \cdot Recall}{\beta^2 Precision + Recall} = \frac{(1 + \beta^2) tp}{(1 + \beta^2)\, tp + \beta^2 \cdot fn + fp},$$

где коэффициент $\beta \in [0; +\infty)$ может рассматриваться как относительная степень важности показателей полноты и точности. При $\beta = 1/2$ точность вдвое важнее полноты, при $\beta = 2$ полнота вдвое важнее точности.

В случае задачи валидации ответов показатель точности является более важным, чем показатель полноты: задача вопросно-ответного поиска разбивалась на две крупные фазы — генерацию гипотез-ответов и проверку гипотез-ответов. Для первой фазы важным показателем качества являлась полнота, для второй — точность [5].

Если рассмотреть модуль валидации как фильтр неверных ответов, то задачей этого фильтра является уменьшение числа ошибок первого рода (fp), может быть даже в ущерб сокращению ошибок второго рода (fn). Чтобы правильно сбалансировать эти два показателя, разложим погрешность $(Error)$ на сумму двух составляющих, соответствующих ошибкам первого и второго рода:

$$Error = \frac{fp + fn}{tp + tn + fp + fn} = Error_I + Error_{\text{II}};$$

$$Error_{\text{I}} = \frac{fp}{tp + tn + fp + fn}, \qquad Error_{\text{II}} = \frac{fn}{tp + tn + fp + fn}.$$

Чтобы подчеркнуть важность ошибок первого рода для задачи валидации, можно определить взвешенную погрешность, в которой ошибки первого и второго рода будут иметь разные веса:

$$E_\alpha = \frac{\dfrac{\alpha \cdot fp + fn}{\alpha + 1}}{tp + tn + \dfrac{\alpha \cdot fp + fn}{\alpha + 1}} = \frac{\alpha \cdot fp + fn}{(\alpha + 1) \cdot (tp + tn) + \alpha \cdot fp + fn};$$

здесь коэффициент $\alpha \in [0; +\infty)$ имеет тот же смысл, что и β в F-мере — относительная степень важности ошибок первого и второго рода. При $\alpha = 1/2$ ошибки второго рода вдвое важнее (менее

желательны) ошибок первого рода, при $\alpha = 2$ ошибки первого рода вдвое важнее ошибок второго рода.

Главным отличием предложенной взвешенной погрешности E_α от F-меры является учет вклада *true-negative* — числа правильно отсеянных ответов. При $tn \gg \max(tp, fn, fp)$, т.е. когда коллекция состоит в основном из отрицательных примеров и фильтр срабатывает правильно, F-мера не меняется, в то время как показатель погрешности стремится к нулю.

Отметим, что тестовая коллекция для валидации ответов действительно должна состоять из преимущественно отрицательных примеров, так как в реальной вопросно-ответной системе на этапе генерации гипотез порождается множество ложных ответов и малое число правильных.

Результаты экспериментов. Предложенный показатель $E_{\alpha=2,0}$ предлагается для сравнения разных алгоритмов валидации ответов[1]. Но так как этот показатель не является общепринятым и вводится впервые, то для каждого эксперимента будем также указывать традиционный показатель $F_{\beta=0,5}$. В табл. 3 представлены результаты экспериментальных прогонов различных алгоритмов валидации ответов, основанных на представлении текста в виде деревьев синтактико-семантических зависимостей. Синтаксико-семантический разбор предложений выполнен с помощью библиотеки AOT.Seman.

Таблица 3

Результаты прогонов различных реализаций модуля валидации ответов

Алгоритм валидации	fn	tn	tp	fp	$Accuracy$	$F_{0,5}$	$E_{2,0}$
Отклонять все ответы	35,8	64,2	0	0	65	0	15,19
Допускать все ответы	0	0	35,8	64,2	35	40	55,37
Пересечение множеств слов [7]	20,3	43,9	18,1	17,7	62	48	23,87
Пересечение множеств связей [7]	31,8	59,2	6,1	2,9	68	43	14,75
Совмещение вершин деревьев [8]	19,6	50,0	15,9	14,5	66	51	19,70
Расстояние редактирования [9]	32,3	60,4	3,3	4,0	64	26	17,44
Параллельный обход графов [1,6]	25,6	61,5	10,0	2,9	71	57	12,80
Сопоставление сказуемых [10]	25,9	60,7	9,7	3,7	70	54	13,64

Примечание. Значения метрик указаны в процентах. $Accuracy/F$ — большее значение лучше. E — меньшее значение лучше.

Из табл. 3 следует, что определение лучших прогонов по метрикам $F_{0.5}$ и $E_{2.0}$ согласуется. Однако метрика $E_{2.0}$ делает конкурентоспособным тривиальный алгоритм “Отклонять все ответы”, что позволяет провести нижнюю границу качества (0,1519 для нашей тестовой коллекции), за которую алгоритмы не должны заходить. Так, алгоритмы “Расстояния редактирования” и “Совмещения вершин деревьев” показывают результаты хуже, чем этот тривиальный алгоритм; F-мера не позволяет выполнять такое сравнение.

Выводы. По результатам участия в кампании РОМИП 2010, было принято решение исследовать подзадачу валидации ответа как задачу бинарной классификации. Была предложена новая метрика — взвешенная погрешность E_α, в отличие от традиционной F-меры учитывающая исходы *true-negative*, являющиеся важной категорией ответов для задачи валидации ответов. При использовании тестовой коллекции вопросов и ответов, состоящей из большого числа негативных примеров, метрика E_α позволяет сравнивать алгоритмы с тривиальным прогоном "Отклонять все ответы".

СПИСОК ЛИТЕРАТУРЫ

1. Соловьев А. А. Алгоритмы валидации ответов в задаче вопросно-ответного поиска // Вестник Воронежского гос. ун-та. Сер.: Системный анализ и информационные технологии. – 2011. – № 2. – С. 181–188.
2. Voorhees E. The TREC-8 question answering track report // In Proc. of the Eighth Text REtrieval Conference (TREC 8). – 1999. – P. 77–82.
3. Voorhees E. M. Overview of the TREC 2002 question answering track // In Proc. of the Eleventh Text Retrieval Conference (TREC 2002). – P. 57–67.
4. Voorhees E. M. Overview of the TREC 2004 question answering track // In Proc. of The Thirteenth Text Retrieval Conference (TREC 2004).
5. Pecas A., Hovy E., Forner P., Rodrigo A., Sutcliffe R., Forascu C. and Sporleder C. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation // Working Notes for the CLEF 2011 Workshop. – 2011.
6. Соловьев А. А. Кто виноват и где собака зарыта? Метод валидации ответов на основе неточного сравнения семантических графов в вопросно-ответной системе // Российский семинар по оценке методов информационного поиска: Тр. РОМИП 2010. (Казань, 15 октября 2010 г.).
7. Wang, Neumann. Using recognizing textual entailment as a core engine for answer validation // Working Notes for the CLEF 2008 Workshop. – 2008.
8. Marsi E., Krahmer E., Bosma W. E., Theune M. Normalized alignment of dependency trees for detecting textual entailment // Second PASCAL Recognising Textual Entailment Challenge. – 10–12 April 2006. – Venice, Italy.
9. Punyakanok V., Roth D. and Yih W. Natural language interface via dependency tree mapping: An application to question answering // AI and Math. – January, 2004.
10. Schlaefer N. A semantic approach to question answering. Saarbrücken 2007.

Статья поступила в редакцию 22.11.2012

Александр Александрович Соловьев — программист научно-технической библиотеки МГТУ им. Н.Э. Баумана, инженер-конструктор ООО "Аплана международные проекты". Автор шести научных работ в области информационного поиска.

A.A. Soloviev — programmer of the scientific and technical library of the Bauman Moscow State Technical University, engineer-designer of OOO APLANA International Projects. Author of 6 publications in the field of data retrieval.