ОБ ОЦЕНКЕ ЭФФЕКТИВНОСТИ СЕМАНТИЧЕСКОГО ХЭЛПЕРА ЭЛЕКТРОННОГО УЧЕБНИКА

Герман Ю. О.

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь, juliagerman@tut.by

Аннотация. Обсуждается стратегический технологический аспект современного дистанционного образования.

Важной задачей построения системы помощи электронного учебника является построение метрики для измерения степени схожести слов. Хорошо известны метрики Левенштейна, Левенштейна-Дамерау, Жаккарда и др. Эти метрики не принимают во внимание смысловую основу слова, которая передается «каркасом» слова, составленным из согласных букв. Для учета указанного обстоятельства рассматриваем два слова – искомое (оригинальное), а также «*скелетон*» искомого слова, из которого выброшены гласные буквы. Для обоих вариантов слов производим разбиение на биграммы и вычисляем оценку сходства с эталоном по формуле

$$\alpha = \frac{n_{1,2}}{n_{search\_pattern}},$$

где $n_{1,2}$ есть число совпадающих биграмм в списках биграмм искомого и эталонного слов, $n_{search\_pattern}$ – размер списка биграмм искомого слова.

Кроме этого, учитываем общее число совпавших букв в заданном слове и эталонном слове. Эта оценка вычисляется таким образом:

$$\beta \equiv \sqrt{q_1 \cdot q_2},$$

где q_1 – процент совпавших букв в слове эталоне; q_2 – процент совпавших букв заданном слове.

Результирующая оценка степени сходства слов выполняется по формуле

$$\alpha = \lambda_1 \cdot \alpha_1 + \lambda_2 \cdot \alpha_2 + \lambda_3 \cdot \beta,$$

$$\lambda_1 + \lambda_2 + \lambda_3 = 1, \quad \lambda_1, \lambda_2, \lambda_3 \geq 0.$$

где λ_1, λ_2, λ_3 определяют приоритеты (веса) для критериев совпадения исходного слова с эталоном (α_1) и скелетона искомого слова со скелетоном эталона (α_2).

Для практического применения экспериментально установлено, что наиболее эффективное распознавание достигается для $\lambda_1 = 0.24,\ \lambda_2 = 0.36, \lambda_3 = 0.4$. Слово считается распознанным, если оценка α не ниже 0.5 (как это принято в теории принятия решений при использовании функций полезности [1]).

Приводится алгоритм поиска искаженного слова на B-дереве специального вида, в узлах которого размещены ассоциированные с понятиями предметной области электронного учебника хэш-коды близких по написанию (лексике) слов.

Литература

1. German O.V., Gourine N.I., Strigalev L.S., German Yu.O. New accents in distant learning. //Дистанционное обучение – образовательная среда XXI века: Материалы VII Междунар.научн-метод. Конференции, 1-2 декаб. 2011 г. – Минск: БГУИР, 2011 – С.300-301.