ЧИСЛЕННЫЕ МЕРЫ “СПЛОЧЕННОСТИ” ИМЕННЫХ ГРУПП

Леонид Леоненко

***Аннотация**: Обсуждаются понятия, функции и алгоритмы, связанные с численными оценками подобия (или “аналогичности”) текстов натурального языка. В так называемой “теории подобия конечных последовательностей” (ТПКП) подобие двух текстов оценивается посредством длины максимальной общей для этих текстов подпоследовательности суб-текстов (например, слов, предложений, etc.) Если в сравниваемых текстах различные вхождения суб-текстов имеют разную значимость сравнительно с другими вхождениями, при оценке степени подобия учитывается не длина, а суммарный “вес” общей для них подпоследовательности суб-текстов (например, суммарный вес общих вхождений слов).*

В данной статье основное внимание уделено мерам оценки структурного сходства текстов. Предполагается, что в тексте-“образце” его суб-тексты сгруппированы в “блоки” (именные группы, предложения, etc.). Принимается следующий принцип “сплоченности”: имена, образующие блок, обычно соседствуют в тексте; и перестановки имен внутри блока разрушают структуру текста в меньшей степени, чем чередование имен, принадлежащих разным блокам.

В статье формулируются понятия, численные меры и алгоритмы, оценивающие степени чередования суб-текстов, принадлежащих разным блокам (меры “сплоченности” блоков). В сочетании с мерами, учитывающими лексическое сходство двух текстов и относительную значимость их сходных суб-текстов, меры сплоченности позволяют адекватно оценивать подобие текстов. Статистические эксперименты показали, что методы ТПКП эффективны в областях автоматической проверки орфографии, идентификации сообщений в телекоммуникационных сетях, компьютерного тестирования знаний. Так, в последней области для компьютера оказывается возможным игнорировать несущественные ошибки в ответах тестируемых, учитывать сокращения и синонимы, разрешить, запретить или ограничить перестановки слов, и т.п.

***Ключевые слова**: аналогия, подобие текстов, алгоритмы оценки подобия, тестирование знаний*

***ACM Classification Keywords**: I.2.6 Artificial Intelligence – Learning – Analogies; K.3.1 Computers and Education - Computer Uses in Education*

Введение

Методы оценки подобия последовательностей составляют важный раздел современной информатики (см., напр., [Смит, 2006]). В работах [Леоненко и Поддубный, 1996], [Leonenko, 2002], [Leonenko, 2010] рассматривались различные численные меры подобия конечных последовательностей и приложения этих мер в задачах компьютерного тестирования знаний. Использование мер подобия позволяет оценивать ответы, вводимые в свободной форме на естественном (русском, украинском, английском и др.) языке.

В упомянутых работах рассматривались меры подобия, связанные главным образом с 1) «элементным» составом последовательностей (в частности, для текстов – с их лексическим составом); и 2) с учетом (неучетом) *линейного порядка* элементов в составе последовательности.

Очевидно, что при оценке подобия предложений и текстов необходимо учитывать не только их лексический состав, но и структуру. В этой статье я рассмотрю численные меры подобия,

предполагающие более «тонкую» структуру текста, чем просто линейный порядок слов. А именно, будут предложены алгоритмы, оценивающие степень чередования, "перемешивания" слов из разных "блоков" текста (именных групп, предложений, etc.).

В первом приближении под «текстом», как и в работе [Леоненко и Поддубный, 1996], понимается иерархически организованная система элементов произвольной природы, для которой значимым является линейный порядок на каждом уровне иерархии. Но теперь он будет *не единственно* значимым отношением на указанном уровне.

Будем считать, что в тексте по каким-либо основаниям выделяются особые ***группы имен*** (например, в русском языке это могут быть группа подлежащего и группа сказуемого в предложении; в алгоритмическом языке – группы, относящиеся к различным операторам программы, etc.). Принимается (сравн.: [Гладкий, 1973], [Добров и др., 2004]) следующий ***«принцип сплоченности» для групп***:

CP *(Cohesion Principle): Имена, образующие группу, обычно соседствуют в тексте; и перестановки имен внутри групп «разрушают» структуру текста в меньшей степени, чем чередование имен, принадлежащих разным группам*.

Проблема, как именно выделять такие группы имен в тексте (в предложении натурального языка, или в конструкции языка программирования, или в формуле, etc.) обсуждается в [Леоненко, 2008].

Ниже предполагается, что заданы два текста **A** и **B**, и в тексте **A** выделены «сплоченные» группы имен. (Когда речь идет о компьютерном тестировании знаний, в качестве **A** выступает *правильный* ответ на тестовый вопрос, а в качестве **B** – *фактический* ответ учащегося). Необходимо оценить подобие, или «аналогичность», текстов **A** и **B**. При этом текст **A** – я буду называть *моделью*, а текст **B** – *прототипом* аналогии (следуя терминологии, принятой в [Уемов, 1971]). Предлагаемые ниже численные меры аналогичности могут применяться к произвольным текстам описанной выше структуры, *независимо* от способа выделения групп имен в тексте.

Меры "разобщенности" элементов заданной группы в тексте

Пусть даны два текста **A** и **B** – модель и прототип; и для модели **A** задано множество групп имен $\{\mathbf{G}_i\}$ $i \in [1,n]$. Говоря неформально, для оценки подобия **A** и **B** будут проверяться следующие главные условия:

1) имеют ли **A** и **B** сходный лексический состав;
2) в какой степени группы имен прототипа **B**, соответствующие группам $\{\mathbf{G}_i\}$ из **A**, «засорены» чужеродными именами.

Аналогично понятию «группы имен в тексте» можно ввести понятие «группы символов в слове» (например, группы букв, относящихся к корню слова в этнических языках, etc.). Всюду ниже термины "символ", "буква" и "слово" будут использоваться как *относительные*: "буква" может пониматься как слово в "слове следующего уровня" (т.е. в предложении), и.т.п.

Численные меры, оценивающие условие 2, базируются на следующем понятии расстояния между символами *a* и *b* из группы $\mathbf{G}_i$ в слове **W**:

Определение 1. Мерой “разобщенности” символов *a* и *b* в слове **W** называется число $\rho(a,b)$ элементов e_k, расположенных в **W** между *a* и *b*, таких что $e_k \in \mathbf{U} \setminus \mathbf{G}_0 \setminus \mathbf{G}_i$.

Здесь **U** – некоторое «универсальное» множество символов, а $\mathbf{G}_0$ – группа «нейтральных» символов, не влияющих на разобщенность.

Теорема 1. Мера разобщенности $\rho(a,b)$ является расстоянием на подмножестве элементов множества $\mathbf{G}_i$, входящих в слово **W**.

Пусть $\mathbf{G}_i[\mathbf{W}]$ – последовательность всех символов группы $\mathbf{G}_i$ в слове **W**. Пусть длина $\mathbf{G}_i[\mathbf{W}]$ равна $\nu \geq 1$, а **M** – матрица размерности $\nu \times \nu$, элементы которой – разобщенности $\rho(a,b)$ всех элементов множества $\mathbf{G}_i[\mathbf{W}]$.

Определение 2. Средней разобщенностью $\mu(\mathbf{G}_i[\mathbf{W}])$ множества $\mathbf{G}_i[\mathbf{W}]$ называется число 0 при $\boldsymbol{\nu}=1$, а при $\boldsymbol{\nu}>1$ – число:

$$\mu(\mathbf{G}_i[\mathbf{W}]) = \frac{\sum_{q=1}^{\nu}\sum_{r=1}^{\nu} M_{q,r}}{\nu(\nu-1)}$$

Определение 3. Спектром разобщенности элементов группы $\mathbf{G}_i$ в слове **W** назовем последовательность натуральных чисел $<L_1, \ldots, L_{\nu-1}>$, где $L_k=\rho(a_k, a_{k+1})$, $k \in [1, \nu-1]$.

Теорема 2. $$\mu(\mathbf{G}_i[\mathbf{W}]) = \frac{2}{v(v-1)} \sum_{j=1}^{v-1} j(v-j)L_j$$

Можно получить различные оценки зависимости меры $\mu(\mathbf{G}_i[\mathbf{W}])$ от длины ν последовательности $\mathbf{G}_i[\mathbf{W}]$, свойств спектра разобщенности, и других параметров. Простейшими являются оценки:

Оценка 1. Пусть $L(\mathbf{G}_i[\mathbf{W}])$ – спектр разобщенности, и пусть *minL* и *maxL* – соответственно минимальный и максимальный элементы $L(\mathbf{G}_i[\mathbf{W}])$. Тогда $minL \times \frac{v+1}{3} \leq \mu(\mathbf{G}_i[\mathbf{W}]) \leq maxL \times \frac{v+1}{3}$.

Оценка 2. Обозначим через *p* сумму элементов спектра $L(\mathbf{G}_i[\mathbf{W}])$, т.е. число всех «чужеродных» для $\mathbf{G}_i$ символов между a_1 и a_v в слове **W**. Тогда $\frac{2p}{v} \leq \mu(\mathbf{G}_i[\mathbf{W}]) \leq \frac{pv}{2(v-1)}$.

Эти оценки можно улучшать, если принимать те или иные гипотезы о структуре разобщающей a_1 и a_v подпоследовательности «чужеродных» для $\mathbf{G}_i$ символов.

Циклические "разобщенности"

Кроме меры $\mu(\mathbf{G}_i[\mathbf{W}])$, можно ввести меру $\mu_O(\mathbf{G}_i[\mathbf{W}])$ так называемой «циклической» разобщенности символов *a* и *b* из группы $\mathbf{G}_i$ в слове **W**. Неформально циклическая разобщенность – это «обычная» разобщенность букв в слове, полученном из **W** «сшиванием» его начала и конца (или «скручиванием **W** в кольцо»). Циклическая разобщенность применяется, главным образом, не к словам, а к предложениям и текстам (и позволяет учесть, например, те случаи, когда в предложении русского языка возможен перенос слов из начала в конец с сохранением смысла).

Пусть множества **U**, $\mathbf{G}_i$ и последовательность **W** определены как в предыдущем разделе; и пусть $\rho(a, b)$ – "обычная" разобщенность букв *a* и *b* из множества $\mathbf{G}_i$ в слове **W**. Пусть, далее, α, β и γ – посдслова слова **W** (возможно, пустые), такие, что

$$\mathbf{W} = < \alpha\, a\, \gamma\, b\, \beta >$$

Пусть, наконец, α' и β' – инверсии слов α и β ; $\mathbf{W}'$– слово, равное

$$\mathbf{W}' = < a\, \alpha'\, \beta'\, b >$$

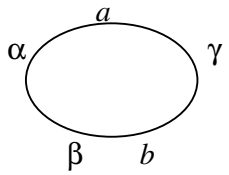
и $\rho'(a,b)$ – "обычная" разобщенность букв *a* и *b* в слове $\mathbf{W}'$.

Определение 4. Мерой циклической разобщенности букв *a* и *b* в слове **W** будем называть число

$$\rho_O(a,b) = \min\{ \rho(a,b) , \rho'(a,b) \}$$

Например, если $c \notin \mathbf{G}_i$, $c \notin \mathbf{G}_0$, $d \in \mathbf{G}_0$, то циклическая разобщенность a и b из $\mathbf{G}_i$ в словах $\mathbf{W}_1 = <acdb>$ и $\mathbf{W}_2 = <accb>$ равна 0 (так как здесь подслова α и β пусты); в словах $\mathbf{W}_3 = <acbc>$ и $\mathbf{W}_4 = <cacb>$ она равна 1; а в слове $\mathbf{W}_5 = <caccbc>$ – равна 2.

Неформальным оправданием термина “циклический” и определения 4 служит схема



из которой видно, в каком смысле инверсии слов α и β располагаются, подобно слову γ, “между” буквами a и b. Формально уместность указанного термина демонстрируется теоремой 3 (см. ниже).

Определение 5. Пусть $\mathbf{W} = < e_1, \dots, e_i, a_1, \dots, a_\nu, e_j, \dots, e_n >$, где $a_1 \in \mathbf{G}_i$, $a_\nu \in \mathbf{G}_i$, причем между a_1 и a_ν расположены все элементы $\mathbf{G}_i$, входящие в слово $\mathbf{W}$. Пусть, далее, $\mathbf{W}_1$ – слово, полученное из $\mathbf{W}$ отбрасыванием элементов $e_1, \dots, e_i$ и $e_j, \dots, e_n$; а знак $\oplus$ обозначает оператор конкатенации последовательностей. *Спектром циклической разобщенности* элементов множества $\mathbf{G}_i$ в слове $\mathbf{W}$ назовем последовательность натуральных чисел

$$\mathrm{Lo}_{\mathbf{G}_i[\mathbf{W}]} = \{L_0\} \oplus \mathrm{L}_{\mathbf{G}_i[\mathbf{W}_1]} \oplus \{L_\nu\} = < L_0 , L_1 , \dots , L_{\nu-1} , L_\nu >$$

где $\mathrm{L}_{\mathbf{G}_i[\mathbf{W}_1]}$ – спектр “обычной” разобщенности для $\mathbf{G}_i$ в слове $\mathbf{W}_1$; L_0 и L_ν – количества тех букв e_k среди соответственно $e_1, \dots, e_i$ и $e_j, \dots, e_n$, которые не принадлежат множеству $\mathbf{G}_0$.

Определение 6. В случае циклической разобщенности *длиной P разобщающей последовательности* для множества $\mathbf{G}_i$ в слове $\mathbf{W}$ будем называть сумму элементов спектра циклической разобщенности $\mathrm{Lo}_{\mathbf{G}_i[\mathbf{W}]}$.

Очевидно, что если p – сумма элементов спектра обычной разобщенности, то $P = p + L_0 + L_\nu$.Следующая лемма также очевидна:

Лемма 1. Если $\mathrm{Lo}_{\mathbf{G}_i[\mathbf{W}]} = < L_0 , L_1 , \dots , L_{\nu-1} , L_\nu >$ – спектр циклической разобщенности, то для любых натуральных $k \geq 1$ и $t \in [k+1, \nu-k]$

$$\rho_O(a_k, a_t) = \min\{ \sum_{j=k}^{t-1} L_j \ , \ \sum_{j=0}^{k-1} L_j + \sum_{j=t+1}^{\nu} L_j \ \} = \min\{ \rho(a_k, a_t), \ P - \rho(a_k, a_t) \}$$

Оценка 3 (следствие леммы 1). При принятых выше обозначениях $\rho_O(a_k, a_t) \leq 0{,}5 * P$.

Теорема 3. Пусть слово $\mathbf{W}'$ – циклическая перестановка слова $\mathbf{W}$, т.е. $\mathbf{W} = < e_1,\dots, e_k, e_{k+1},\dots, e_n >$ и $\mathbf{W}' = < e_{k+1},\dots, e_n, e_1,\dots, e_k >$. Пусть, далее, вхождения a и b некоторых букв слова $\mathbf{W}$ принадлежат множеству $\mathbf{G}_i$, а a' и b' – вхождения в слово $\mathbf{W}'$, в которые перешли a и b в результате циклической перестановки $\mathbf{W}$. Тогда $\rho_O(a', b') = \rho_O(a , b)$.

Теорема 4. Мера циклической разобщенности $\rho_O(a,b)$ является расстоянием на подмножестве элементов множества $\mathbf{G}_i$, входящих в последовательность $\mathbf{W}$.

Определение 7. Пусть $\nu > 1$. *Матрицей циклической разобщенности* множества $\mathbf{G}_i[\mathbf{W}] = \{a_1, \dots, a_\nu\}$ назовем матрицу $\mathbf{M}^{\mathbf{O}}$ размерности $\nu \times \nu$, элементы которой $\mathbf{M}^{\mathbf{O}}_{q,r}$ суть $\rho_O(a_q, a_r)$ в последовательности $\mathbf{W}$.

Определение 8. *Средней циклической разобщенностью* $\mu_O(\mathbf{G}_i[\mathbf{W}])$ множества $\mathbf{G}_i[\mathbf{W}]$ назовем число 0 при ν=1, а при ν>1 – число:

$$\mu_O(\mathbf{G}_i[\mathbf{W}]) = \frac{\sum_{q=1}^{\nu}\sum_{r=1}^{\nu} M^{\circ}_{q,r}}{\nu(\nu-1)}$$

где $\mathbf{M}^{\mathbf{O}}$ – матрица циклической разобщенности множества $\mathbf{G}_i[\mathbf{W}]$.

Следствием теоремы 4 является равенство (при ν>1)

$$\mu_O(\mathbf{G}_i[\mathbf{W}]) = 2\times\frac{\sum_{q=1}^{\nu-1}\sum_{r=q+1}^{\nu} M^{\circ}_{q,r}}{\nu(\nu-1)}$$

Из теоремы 3 следует: если слово $\mathbf{W'}$ – циклическая перестановка слова $\mathbf{W}$, то $\mu_O(\mathbf{G}_i[\mathbf{W}]) = \mu_O(\mathbf{G}_i[\mathbf{W'}])$.

Теорема 5. Для любого слова $\mathbf{W}$ имеет место неравенство $\mu_O(\mathbf{G}_i[\mathbf{W}]) \le \mu(\mathbf{G}_i[\mathbf{W}])$.

Оценка 4. Пусть *min*L и *max*L – соответственно минимальный и максимальный элементы спектра циклической разобщенности $L\varnothing_{\mathbf{G}_i[\mathbf{W}]}$. Тогда

$$min\mathrm{L}\times\frac{\nu+1}{4} \le \mu_O(\mathbf{G}_i[\mathbf{W}]) \le max\mathrm{L}\times\frac{1}{4}(\nu+1+\frac{1}{\nu-1})$$

Оценки 3 и 4 можно улучшать, если принимать те или иные гипотезы о структуре спектра циклической разобщенности.

Понятие "φ–сплоченности"

Определение 9. Пусть φ – любая функция со следующими свойствами: 1) $\varphi(x)$ определена для любого $x \ge 0$; 2) $\varphi(0) = 1$; 3) $\varphi(x)$ невозрастает при $x \to \infty$. Тогда *φ–сплоченностью* элементов множества $\mathbf{G}_i$ в слове $\mathbf{W}$ назовем величину

$$C(\varphi, \mathbf{G}_i, \mathbf{W}) = \varphi(\mu(\mathbf{G}_i[\mathbf{W}]))\ ,$$

а *циклической φ–сплоченностью* – величину

$$C_O(\varphi, \mathbf{G}_i, \mathbf{W}) = \varphi(\mu_O(\mathbf{G}_i[\mathbf{W}]))$$

Ввиду свойства 3 функции φ и теоремы 5 предыдущего раздела $C_O(\varphi, \mathbf{G}_i, \mathbf{W}) \ge C(\varphi, \mathbf{G}_i, \mathbf{W})$.

Если трактовать слово $\mathbf{W}$ как *предложение* (или иное корректное выражение) некоторого (натурального или искусственного) языка; буквы $\mathbf{W}$ как *слова* этого выражения; а множества $\mathbf{G}_i$ – как некоторые *части* $\mathbf{W}$ (например, именные группы, входящие в $\mathbf{W}$); то средние разобщенности μ и μ_O являются оценками "сплошности" размещения слов части $\mathbf{G}_i$ в выражении $\mathbf{W}$. Циклическая либо нециклическая разобщенность используется в зависимости от того, допускает ли ситуация перенос части слов $\mathbf{G}_i$ из начала в конец выражения без ущерба для "сплошности" $\mathbf{G}_i$. Например, для следующего предложения (из «Ночи перед Рождеством» Н.В. Гоголя):

$\mathbf{W}$ = *Ведьма, увидевши себя вдруг в темноте, вскрикнула*

такие переносы допустимы, но лишь ограниченно: предложение

Вскрикнула ведьма, увидевши себя вдруг в темноте

вероятно, можно считать эквивалентным **W**; однако

В темноте вскрикнула ведьма, увидевши себя вдруг

– уже нет.

Тем не менее, иногда разумно применять именно циклическую меру разобщенности. Так, в работе [*Добров и др.*, 2004, с.66] обсуждается проблема распознавания в техническом тексте понятий, описанных формальной онтологией соответствующей предметной области. Авторы пишут: «распознавание "разорванных" терминов в тексте представляет достаточно сложную проблему», и приводят следующий пример: «...для понятия *вертикальный маневр* существует синоним *маневр в вертикальной плоскости*, который мог бы встретиться в предложении вида "*Маневр осуществляется самолетом... в вертикальной плоскости*"». Очевидно, что здесь циклическая сплоченность упомянутого термина, в отличие от нециклической, не нарушена.

Что касается φ–сплоченности, то она будет использоваться как особая мера сходства таких последовательностей слов из $\mathbf{G}_i$, которые в различной степени "засорены чужеродными словами". Этот тип сходства можно сопоставить, не претендуя на полноту аналогии, с понятием "узнаваемость". Условие 2), которому удовлетворяет φ, означает, что когда между словами из $\mathbf{G}_i$ в **W** нет слов, принадлежащих другим частям **W** (кроме "нейтральных" слов из $\mathbf{G}_0$), узнаваемость части $\mathbf{G}_i$ в **W** равна 1. В противном случае она может быть (не обязательно) меньше 1. Конкретный вид функции φ может выбираться в зависимости от принимаемой гипотезы, насколько узнаваемость части $\mathbf{G}_i$ искажается от ее "засорения". Если, допустим, принимается, что даже одно "разобщающее" слово сильно портит узнаваемость $\mathbf{G}_i$, рационально выбрать $\varphi = e^{-x}$. Если же считать, что $\mathbf{G}_i$ остается хотя бы "наполовину" узнаваемым, сколько бы чужеродных слов его не перемежали, можно выбрать $\varphi = (1+x)/(1+2x)$, пределом которой при $x \to \infty$ будет 1/2. Наконец, если принять, что при любой степени "засорения" узнаваемость снижается ровно вдвое, можно положить $\varphi(x)=1$ при $x=0$ и $\varphi(x) = 0{,}5$ при $x>0$.

О применении мер φ–сплоченности

Приведем примеры применения рассмотренных мер к оценкам аналогичности структур различных предложений русского языка, сходных по лексическому составу. Предметная область – компьютерное тестирование знаний. Пусть «эталонным» ответом на вопрос

Каково наиболее значительное достижение в области математики конца XVII в.; и кто его автор?

считается предложение

Ньютон и Лейбниц открыли математический анализ.

Это предложение трактуется как модель вывода по аналогии; и в нем выделяются именные группы: $\mathbf{G}_1$ – множество слов, подобных одному из слов {*Ньютон*, *Лейбниц*}; $\mathbf{G}_2$ – множество слов, подобных для {*изобрел*, *открыл*, *придумал*, *создал*, *автор*}; $\mathbf{G}_3$ – множество слов, подобных для {*математический*, *анализ*, *матанализ*}. Подобие трактуется в смысле [Леоненко и Поддубный, 1996]: задается некий численный уровень подобия, достаточный, чтобы, например, слова *Ляйбниц* или *Лейбницем* считались эквивалентными слову *Лейбниц*.

Пусть в качестве других возможных ответов на поставленный вопрос компьютер должен оценить предложения:

1. *Лейбниц и Ньютон – изобретатели матанализа*.
2. *Мат. анализ был создан Ньютоном и Лейбницем*.
3. *Матанализ придуман Лейбницем и Ньютоном*.

4. *Лейбниц, а также Ньютон, открыли анализ*.
5. *Ньютон открыл матанализ, и Лейбниц также*.
6. *Мат. анализ Ньютон открыл, и Лейбниц*.
7. *Матанализ, Лейбниц, Ньютон*.
8. *Лейбниц придумал матанализ*.
9. *Анализ Ньютона математический Лейбниц придумал*.
10. *Математика Ньютона открыла Лейбницев анализ*.
11. *Ньютонов анализ создал математику Лейбница*.
12. *Открыт и Ньютон анализом Лейбница*.
13. *Ньютон придумал Лейбница и создал анализ*.
14. *Анализ изобретений Ньютона и Лейбница открыл мат*.
15. *Ньютон открыл математику, Лейбниц – анализ* '*и*'.

Обозначим эти шестнадцать предложений через $\mathbf{W}_0$ (эталон), $\mathbf{W}_1$,…, $\mathbf{W}_{15}$. Видимо, не встретит возражений утверждение, что $\mathbf{W}_1$ – $\mathbf{W}_5$ являются приемлемыми ответами. Ответ $\mathbf{W}_6$ стилистически неудачен, но верен. Ответ $\mathbf{W}_7$ предельно лапидарен и верен. Ответ $\mathbf{W}_8$ неполон; можно спорить, следует ли "засчитывать" его. При устном опросе преподаватель, возможно, тем или иным способом побудил бы студента дополнить ответ $\mathbf{W}_8$. Мы будем считать $\mathbf{W}_8$ "пограничным" между допустимыми и недопустимыми ответами. Что касается ответа $\mathbf{W}_9$, то его можно истолковать как двусмысленный (или, если угодно, косноязычный) – и потому также отнести к "пограничным" ответам. Либеральный экзаменатор примет его, ригорист – нет (но и ригорист задумается, если ответ $\mathbf{W}_9$ поступил от иностранца, плохо владеющего русским).

Ответы же $\mathbf{W}_{10}$ – $\mathbf{W}_{15}$, по-видимому, являются совершенно неприемлемыми. Но все же "качество" их неприемлемости разное. Ответ $\mathbf{W}_{10}$ выражает – правильным русским языком – вполне ложное суждение. Но представим, что его грамматическая правильность лишь кажущаяся, поскольку $\mathbf{W}_{10}$ (как выше $\mathbf{W}_9$) высказан иностранцем, не владеющим падежами. В этой ситуации $\mathbf{W}_{10}$ становится весьма похожим на $\mathbf{W}_9$, хотя все же менее приемлемым.

Постоянная "гипотеза иностранца" может показаться нарочитой и неоправданной. Однако ясно, что те или иные искажения могут появиться в содержательно верном ответе по разным причинам. Представим себе, например, вполне русскоязычного студента, который переиначивает слова в ответе, желая посмеяться над компьютерной тестирующей системой. Если конструкторы тестирующей системы хотят предусмотреть какую-то реакцию на подобные действия студентов, они могут зафиксировать степень допустимых искажений ответов. Теория подобия текстов [Леоненко и Поддубный, 1996] позволяет задать пределы *буквенных* искажений слов фразы. В нашем же случае мы хотим учесть те искажения, которые вносятся *перемещением слов, принадлежащих одним именным группам, внутрь других именных групп* предложения-ответа.

Допустим поэтому, что: (**A1**) Во всех рассматриваемых предложениях $\mathbf{W}_0$ – $\mathbf{W}_{15}$ падежные окончания, а также иные "незначительные" искажения слов (вроде *Лейбниц* – *Ляйбниц*) не следует принимать во внимание; (**A2**) Оттенки смысла слов-синонимов вроде *открыл*, *изобрел* и *создал* (эти оттенки в той или иной мере существенны при трактовке $\mathbf{W}_0$ – $\mathbf{W}_{15}$ как грамматически правильных фраз русского языка), также игнорируются. Наконец, учтем, что при вводе ответа с клавиатуры некоторые буквы слов могут быть непреднамеренно пропущены – и это же относится к коротким, особенно однобуквенным, словам. Поэтому примем: (**A3**) Пропуск/добавление одного-двух однобуквенных слов не должны сильно влиять на оценку ответа.

Ясно, что допущения (**A1**)–(**A3**) не следует трактовать как "общезначимые". Вполне уместны (и неоднократно применялись) подходы, при которых, скажем, от тех или иных падежных окончаний слов существенно зависел способ распознавания смысла предложения (см., напр., [Гладун, 1987, с.100-119], [Добров и др., 2004, с.59]). Но при анализе предложения на "сплоченность" его именных групп принятие (**A1**)–(**A3**) оправданно, особенно если применять такой анализ ***независимо*** от других способов оценки предложения. Например, можно ***сначала*** отобрать приемлемый ответ на основе лексического подобия фраз, ***не*** предполагая при этом (**A1**)–(**A3**), и лишь ***затем*** оценить степень сплоченности этого ответа с учетом (**A1**)–(**A3**). Именно это предполагает метод оценки подобия текстов, описанный ниже.

При допущениях (**A1**)–(**A3**) предложение $\mathbf{W}_{11}$: *Ньютонов анализ создал математику Лейбница* – может быть расценено как искаженное *Ньютон анализ создал математический, и Лейбниц*, что почти совпадает с $\mathbf{W}_6$ или $\mathbf{W}_9$.

Таким образом, если принимать (**A1**)–(**A3**), то ответы $\mathbf{W}_{10}$, $\mathbf{W}_{11}$ (и, как легко видеть, $\mathbf{W}_{12}$) нужно оценить как хотя и неприемлемые, но все же находящиеся от множества "пограничных" ответов "на меньшем расстоянии", чем, ответы $\mathbf{W}_{13}$ – $\mathbf{W}_{15}$.

Нетрудно убедиться в том, что мера широкого невзвешенного подобия F_0 [Леоненко и Поддубный, 1996], оценивающая подобие фактического и эталонного ответа без учета порядка следования слов, припишет предложениям $\mathbf{W}_1$, …, $\mathbf{W}_{15}$ следующие степени сходства с $\mathbf{W}_0$ (с точностью до сотых):

$\mathbf{W}_1$ – 1.0; $\mathbf{W}_2$ – 0.86; $\mathbf{W}_3$ – 1.0; $\mathbf{W}_4$ – 1.0; $\mathbf{W}_5$ – 0.83; $\mathbf{W}_6$ – 1.0; $\mathbf{W}_7$ – 0.6;
$\mathbf{W}_8$ – 0.6; $\mathbf{W}_9$ – 0.83;
$\mathbf{W}_{10}$ – 0.83; $\mathbf{W}_{11}$ – 0.83; $\mathbf{W}_{12}$ – 1.0; $\mathbf{W}_{13}$ – 0.83; $\mathbf{W}_{14}$ – 0.86; $\mathbf{W}_{15}$ – 1.0.

Таким образом, мера F_0 не отделяет удовлетворительным образом приемлемые ответы {$\mathbf{W}_1$,…, $\mathbf{W}_7$} от "пограничных" {$\mathbf{W}_8$, $\mathbf{W}_9$} и от неприемлемых {$\mathbf{W}_{10}$,…, $\mathbf{W}_{15}$}.

Ситуация практически не улучшается, если применить взвешенное подобие [Leonenko, 2002], [Leonenko, 2010]. Давайте припишем элементам эталона следующие относительные веса (в скобках):

*Ньютон***(4)** *и***(0)** *Лейбниц***(4)** *открыли***(1)** *математический***(3)** *анализ***(8).**

Заметим, что синонимичные словосочетания могут содержать разное количество слов, но их суммарные веса должны быть равны; таким образом, *матанализ* из группы $\mathbf{G}_3$ будет иметь вес 11, равный сумме весов слов *математический* и *анализ*. Кроме того, при задании весов эталона можно указать множество слов, которые, хотя и не рассматриваются как синонимичные словам эталона, при появлении в ответе должны иметь вес 0 – таковы, для данного примера, слова *также*, *кроме*, *был*, etc.

Далее применим меру широкого (не учитывающего порядок слов) *взвешенного* подобия, которая принимает ожидаемый вес слов ответа, не вошедших в базис подобия, равным *среднему арифметическому слов* эталона, *имеющих ненулевой вес* (в нашем случае этот ожидаемый вес равен 4). Обозначим эту меру через V_0(**A**, **B**), где **A** – текст-модель (эталон), и **B** – текст-прототип. V_0–взвешенные степени подобия пятнадцати предложений-ответов эталону $\mathbf{W}_0$ будут такими:

$\mathbf{W}_1$ – 1.0; $\mathbf{W}_2$ – 1.0; $\mathbf{W}_3$ – 1.0; $\mathbf{W}_4$ – 0.85; $\mathbf{W}_5$ – 1.0; $\mathbf{W}_6$ – 1.0; $\mathbf{W}_7$ – 0.95;
$\mathbf{W}_8$ – 0.8; $\mathbf{W}_9$ – 1.0;
$\mathbf{W}_{10}$ – 1.0; $\mathbf{W}_{11}$ – 1.0; $\mathbf{W}_{12}$ – 0.85; $\mathbf{W}_{13}$ – 0.81; $\mathbf{W}_{14}$ – 0.83; $\mathbf{W}_{15}$ – 1.0.

(Здесь при подсчете степеней для $\mathbf{W}_{13}$ и $\mathbf{W}_{14}$ учтено, что каждое из них содержит *одно* слово – например, для $\mathbf{W}_{14}$ им может считаться любое из слов *изобрел* или *открыл*, а также подобные им слова, – не входящее в базис подобия).

Улучшения, вносимые мерой V_0 сравнительно с F_0, связаны только с относительным повышением веса ответа $\mathbf{W}_7$ по сравнению с $\mathbf{W}_8$. Вместе с тем классы приемлемых, неприемлемых и "пограничных"

ответов по-прежнему не отделены.

Попробуем получить искомое отделение классов, использовав понятие φ–сплоченности. Вычислим нециклические φ–сплоченности каждого из указанных шестнадцати предложений $\mathbf{W}_0$, …, $\mathbf{W}_{15}$, для каждой из выделенных в $\mathbf{W}_0$ именных групп $\mathbf{G}_1$, $\mathbf{G}_2$ и $\mathbf{G}_3$, использовав в качестве φ функцию $\varphi=1/(1+x)$. Мы получим, в частности: $C(\varphi, \mathbf{G}_1, \mathbf{W}_0) = 1$; $C(\varphi, \mathbf{G}_1, \mathbf{W}_1) = 1$; $C(\varphi, \mathbf{G}_1, \mathbf{W}_5) = 0.333$; и т.д.

В теории подобия текстов *базисом подобия* текстов **A** и **B** называется, если опустить детали, то мультимножество (для широкого подобия) или кортеж (для узкого подобия) их общих вхождений субтекстов, на основании которого оценивается степень подобия **A** и **B**. Для оценки невзвешенного подобия важен размер (мощность) базиса, а для взвешенного – суммарный вес элементов базиса.

Пусть для любого предложения $\mathbf{W}_k$ с $k\in [1,15]$ через $\mathbf{B}(\mathbf{W}_k)$ обозначен *базис широкого взвешенного подобия* фразы $\mathbf{W}_k$ относительно эталона $\mathbf{W}_0$. Скорректируем правило для вычисления степени широкого взвешенного подобия $V_0(\mathbf{W}_0, \mathbf{W}_k)$ следующим образом:

Если в базис подобия $\mathbf{B}(\mathbf{W}_k)$ входит группа слов из множества $\mathbf{G}_i$, то при подсчете веса базиса суммарный вес этой группы следует умножить на φ–сплоченность множества $\mathbf{G}_i$ в предложении $\mathbf{W}_k$.

Численную оценку подобия двух фраз, полученную в результате применения приведенного выше правила, учитывающего φ–сплоченности групп слов во фразе, обозначим через $H_0(\mathbf{A},\mathbf{B})$. Ясно, что в общем случае H_0 зависит от выбранной функции φ и типа сплоченности (циклической либо нециклической).

В нашем примере (нециклическая φ–сплоченность с $\varphi=1/(1+x)$) H_0–взвешенные степени подобия пятнадцати предложений-ответов эталону $\mathbf{W}_0$ окажутся следующими:

$\mathbf{W}_1$ – 1.0; $\mathbf{W}_2$ – 1.0; $\mathbf{W}_3$ – 1.0; $\mathbf{W}_4$ – 0.85; $\mathbf{W}_5$ – 0.73; $\mathbf{W}_6$ – 0.8; $\mathbf{W}_7$ – 0.95;
$\mathbf{W}_8$ – 0.8; $\mathbf{W}_9$ – 0.53;
$\mathbf{W}_{10}$ – 0.39; $\mathbf{W}_{11}$ – 0.43; $\mathbf{W}_{12}$ – 0.65; $\mathbf{W}_{13}$ – 0.63; $\mathbf{W}_{14}$ – 0.53; $\mathbf{W}_{15}$ – 0.46.

Мы видим, что одновременный учет взвешенного лексического подобия и нециклической φ–сплоченности позволяют хорошо отделить классы допустимых {$\mathbf{W}_0$,…, $\mathbf{W}_7$} и недопустимых {$\mathbf{W}_{10}$,…, $\mathbf{W}_{15}$} ответов. Скажем, если выбрать уровень подобия, на котором ответ признается приемлемым, равным 0.7; то все содержательно правильные ответы будут признаны компьютером допустимыми, а содержательно неправильные – недопустимыми. При этом “пограничный” ответ $\mathbf{W}_8$ будет отнесен к классу допустимых, а “пограничный” ответ $\mathbf{W}_9$ признан недопустимым – что, по-видимому, приемлемо.

Если вместо нециклической использовать *циклическую* φ–сплоченность, то для предложения $\mathbf{W}_5$ степень подобия эталону $\mathbf{W}_0$ окажется равной 1. Но нетрудно показать, что в целом для совокупности предложений $\mathbf{W}_0$, …, $\mathbf{W}_{15}$ циклические φ–сплоченности оказываются непригодными для различения классов допустимых и недопустимых ответов в множестве $\mathbf{W}_1$,…, $\mathbf{W}_{15}$ (причем при различных выборах функции φ). Скажем, степени подобия эталону $\mathbf{W}_0$ для ответов $\mathbf{W}_6$ и $\mathbf{W}_{10}$ получаются равными; степень подобия $\mathbf{W}_6$ меньше, чем у $\mathbf{W}_{14}$, и т.п.

Заключение

Принцип сплоченности ***CP*** является одним из структурных принципов, нарушения которого снижают «целостность» восприятия текста человеком. Эта целостность обусловлена семантическими связями терминов текста (как для натуральных, так и для искусственных языков). Тем не менее сам принцип ***CP*** носит чисто синтаксический характер. В этом он сходен с принципом взвешенного лексического подобия.

Проведенные статистические эксперименты [Баранов, 2004] показали, что в определенных областях, – в частности, в сфере компьютерного тестирования знаний, – алгоритмы сравнения текстов, базирующиеся на описанных в данной статье мерах подобия, позволяют получить адекватные оценки подобия текстов. Оценки «открытых» ответов, выставляемые системой тестирования знаний CONTROL (использующей упомянутые алгоритмы), статистически не отличаются от оценок специалистов-преподавателей. Это дает возможность *игнорировать разного типа несущественные ошибки и искажения* ответов; и как следствие позволяет сделать «открытые» ответы равноправными с другими типами ответов при компьютерном тестировании (см. [Леоненко, 2010]).

Благодарности

Работа опубликована при финансовой поддержке проекта **ITHEA XXI** Института информационных теорий и приложений FOI ITHEA Болгария www.ithea.org и Ассоциации создателей и пользователей интеллектуальных систем ADUIS Украина www.aduis.com.ua.

Библиография

[Leonenko, 2002] L. Leonenko. Analogical inferences in computer assisted knowledge testing systems // 6th Multi-Conference on Systemics, Cybernetics and Informatics. – Orlando, Florida, USA, 2002. Proc., Vol. XVIII, pp.371-376.

[Leonenko, 2010] L. Leonenko. Analogies between Texts: Mathematical Models and Applications in Computer-assisted Knowledge Testing // Information Models of Knowledge. – Kiev, Ukraine – Sofia, Bulgaria: ITHEA, 2010, pp. 128 – 134.

[Баранов, 2004] В. Ю. Баранов. Комп'ютерне тестування з інформатики: підсумки педагогічного експерименту в Одеській національній академії зв'язку // Теорія та методика навчання математики, фізики, інформатики. – Вип. 4. – Т. 3 – Кр.Ріг.: НМетАУ, 2004. – с. 6 –12.

[Гладкий, 1973] А.В. Гладкий. Математические методы изучения естественных языков // Труды МИАН им. В.А.Стеклова. – 1973. – Том 133. – с. 95 – 108.

[Гладун, 1987] В.П. Гладун. Планирование решений. – К.: Наукова думка, 1987. – 168 с.

[Добров и др., 2004] Б.В. Добров, Н.В. Лукашевич, О.А. Невзорова, Б.Е. Федунов. Методы и средства автоматизированного проектирования прикладной онтологии // Изв. РАН. Теория и системы управления. – 2004. – № 2. – с. 58–68.

[Леоненко, 2008] Л.Л. Леоненко. Язык тернарного описания в оценках связности текстов // Сб. трудов VIII международной конф. "Интеллектуальный анализ информации". – К.: Просвіта, 2008, с. 286–295.

[Леоненко, 2010b] Л. Леоненко. "Открытые" ответы в компьютерном тестировании знаний // Information Models of Knowledge. – Kiev, Ukraine – Sofia, Bulgaria: ITHEA, 2010, pp. 355 – 361.

[Леоненко и Поддубный, 1996] Л.Л. Леоненко, Г.В. Поддубный. Теория подобия конечных последовательностей и ее приложения к распознаванию образов // Автоматика и телемеханика, 1996, № 8, с.119-131.

[Смит, 2006] У. Смит. Методы и алгоритмы вычислений на строках. – М.: "И.Д. Вильямс", 2006. — 496 с.

[Уемов, 1971] А.И. Уемов. Логические основы метода моделирования. М.: Мысль, 1971. – 311 с.

Информация об авторе

Леонид Леоненко *– Одесская национальная академия связи им. А. С. Попова, доцент, ул. Конная, 22, кв. 6, Одесса, 65029, Украина; e-mail: Leonid.Leonenko@gmail.com*

Основные области научной деятельности: неклассическая логика, общая теория систем, компьютерное тестирование знаний