

РАСПОЗНАВАНИЕ КОМПЛЕКСНОЙ ПАРЫ В КОНЕЧНОМ МНОЖЕСТВЕ МЕТОДАМИ НЕЧЕТКОЙ МАТЕМАТИКИ

Д.т.н. С.А. Нестеренко, О.И. Дадерко, А.В. Торопенко

Одесский национальный политехнический университет

Украина, Одесса

alla.androsyk@gmail.com

Предложено некоторые концепции компьютерной поддержки автоматизированного перевода стихотворного текста. Формализованы основные атрибуты текста – ритм и рифму. Разработан алгоритм автоматизированного подбора рифмуемых слов. Предложены информационные показатели для рифмованной пары.

Ключові слова: стихотворный текст, информационный показатель, автоматизированный перевод.

Предлагаемая система подготовки принятия решений (СПР) при автоматизированной обработке текста основывается на достижениях специалистов в области машинного перевода [1]. При объединении этих достижений возникли проблемы, специфичные именно для машинного перевода стихов.

В частности, переосмысления потребовал даже такой установившийся факт, как количество букв, поскольку для поддержания при автоматизированном переводе необходимого ритма стихотворного текста, необходимо различения системой ударных и безударных гласных, что добавляет к 33 известным русским буквам еще 9 ударных гласных (буква «ё» в русском языке всегда ударная) [2]. Наличие 42 букв потребовало полностью «переписать» словарь русского языка, для чего была создана соответствующая база данных.

Для сохранения соответствия, смысла, образности и новизны текста необходимо было разработать новую концептуальную, восходящую исключительно к проблемам информатики, основу системы, так как обычные переводчики оставляют «за кадром» те особенности текста, без которых не обходится живая речь.

Для разработки *информационной основы* СПР, позволяющей в автоматизированном режиме подбирать необходимые элементы (слова, словосочетания) обрабатываемого текста, художественное произведение рассматривалось с двух сторон – творческой и технической (информационной).

Во всех случаях в СПР машина лишь подбирает необходимые слова и словосочетания (*пары*), окончательный же выбор остается за человеком. В системе имеются также многочисленные настройки пользователя (НП), благодаря которым последний может существенно изменить результат машинной части перевода в необходимую сторону. Все предложенные подсистемы относятся к русскому языку. При переводе на тот или иной язык, отличный от русского, они должны быть предварительно созданы для этого языка.

В основу формализации понятия «пара» положены следующие концепции.

1. В паре участвуют два слова. В СПР первое слово задается, а второе ищется.

2. Группа или часть группы согласных букв, расположенных между двумя гласными, для расчетов информационных показателей пары может объединяться в одну условную согласную букву.

Например, в паре «**устроил** – **беспокоил**» буквы «стр» в слове «**устроил**» могут считаться одной согласной и сравниваться с согласной «к» в слове «**беспокоил**».

3. Согласная буква (группа согласных) в конце слов, входящих в пару, для расчетов информационных показателей пары может опускаться. Например, в паре «**лето** – **рассветов**» буква «в» в конце слова «**рассветов**» при расчетах может не учитываться.

4. Введем следующие информационные показатели пары:

– **общая длина** – количество учитываемых букв (с конца строки), берущихся во внимание при определении перечисленных ниже информационных показателей пары;

– **сила** – комплексное число, характеризующее близость слов, входящих в пару;

– **новизна** – число, характеризующее частоту употребления пары в опубликованных произведениях;

– **эстетичность** – балл, экспертная оценка, характеризующая степень невозможности по тем или иным эстетическим причинам использования этой пары.

Общая длина пары может быть определена «вручную» – прямым усмотрением (например, длина пары «и Тань я – скитанья» может быть оценена специалистом в шесть букв) или вычислена автоматически.

Сила пары – комплексное число, состоящее из двух частей – четкой и нечеткой. Например, пара «художник – придорожных» при длине в шесть букв содержит совпадение букв «ожн», стоящих на позициях 3 – 5 с конца. Именно они участвуют в расчетах четкой части силы пары. Остальные буквы рифмы, стоящие на позициях 1, 2 и 6, принимаются во внимание при расчете нечеткой части.

Четкая часть Ч – суть *количественная* оценка по предложенной формуле меры полного совпадения букв из 42-буквенного алфавита в словах пары на соответствующих местах.

Нечеткая часть Н – суть *количественная* оценка с помощью теории нечетких множеств меры близости несовпадающих букв из 42-буквенного алфавита в словах пары на соответствующих местах.

Сила пары, как и обычное комплексное число, может рассматриваться как векторная сумма **Р** четкого и нечеткого векторов $\mathbf{P} = \mathbf{Ч} + \mathbf{Н}$ и рассчитываться по правилу скаляра:

$$P^2 = Ч^2 + Н^2 + 2ЧН \cdot \cos(\mathbf{Ч}, \mathbf{Н}).$$

Новизна пары – число, рассчитываемое по уравнению $K = 1/k$, где k – количество существующих в банке данных аналогичных рифм.

Эстетичность пары N – число определяемое экспертной оценкой. Пара, обладающая большой силой и высокой новизной, может быть отнесена экспертами к очень неэстетичной и не рекомендована к употреблению по этому параметру. Например, пара «кровь – любовь» сильнее, чем «вновь – любовь», однако новизна ее ниже, а неэстетичность значительно выше.

5. В базе данных СПР хранятся:

– слова русского языка во всех склонениях, спряжениях и числах, составленные в 42-буквенном алфавите; система управления этой базой данных, была построена таким образом, чтобы она позволяла пользователю подбирать слова не по первым буквам алфавита, как в обычных словарях, а по последним – необходимый атрибут при подборе соответствующих пар;

– известные (применявшиеся в опубликованных текстах) пары с указанием их длины, силы, новизны и неэстетичности.

Практическое использование созданной СПР показало хорошие результаты, позволившие рекомендовать систему для применения ее при «ручном» и полуавтоматическом (с предварительным использованием подстрочника) переводе тестов.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1 Становский А.Л., Науменко Е.А., Панченко А.Н. Концептуальная основа автоматизированного переводчика стихотворного текста / Материалы XIV конференции «Моделирование в прикладных научных исследованиях». Одесса: ОНПУ, 2007. С. 78.

2. Становский А.Л., Поздняков В.Ю. Некоторые концепции системы поддержки автоматизированного перевода стихотворного текста / Труды Одесского политехнического университета, 2008.

Nesterenko S. A., Daderko O.I., Toropenko A.V.

Recognition of complex pair in fin multiple methods of fuzzy mathematics

Some concepts of computer support for the automated translation of the verse text are proposed. The basic attributes of the text - rhythm and rhyme - are formalized. An algorithm for the automated selection of rhyme words has been developed. Information indicators for the rhymed pair are proposed..

Keywords: poetic text, information indicator, automated translation