7А

УДК 681.3

А.В. Анисимов, А.А. Марченко
Киевский национальный университет имени Тараса Шевченко, г. Киев, Украина
rozenkrans@yandex.ru, ava@unicyb.kiev.ua

Ассоциативное реферирование естественно-языковых текстов

В статье рассматривается построение ассоциативно-семантического алгоритма реферирования текстов на естественном языке. Ассоциативный анализ и моделирование ассоциативно-семантических связей в тексте открывают широкие перспективы решения фундаментальных проблем компьютерной лингвистики.

В настоящее время в системах искусственного интеллекта в целом и в компьютерной лингвистике в частности, чрезвычайно популярным и актуальным стало направление ассоциативно-семантического анализа в решениях сложных плохо формализируемых нечетких задач. Как известно, человеческий мозг, являющийся субстратом естественного интеллекта, состоит из левого и правого полушарий. Согласно широко распространенной гипотезе, левое полушарие занимается логической и рекурсивной обработкой данных, в то время как правое занято ассоциативным поиском и установлением ассоциативно-семантических связей в структурах данных мозга [1]. Типичными примерами проявления работы левого полушария мозга человека являются решение математических и логических задач, игры переборного типа, изучение точных прикладных наук. Функции правого полушария проявляются в процессе творческой, гуманитарной деятельности, ассоциативного мышления, интуитивной работы мозга. Классические направления искусственного интеллекта всегда уделяли и уделяют первоочередное внимание изучению и развитию методов и алгоритмов, моделирующих «левостороннюю» логико-рекурсивную работу мозга. В то же время функции правого полушария до сих пор остаются в тени. Во многом так сложилось из-за трудности понимания и восприятия таких явлений и процессов, как ассоциативное мышление, интуиция, творчество. Но именно односторонний подход к решению фундаментальных задач искусственного интеллекта и является его «ахиллесовой пятой», делающей большинство проблем данного направления кибернетики принципиально неразрешимыми. На примере задач анализа и синтеза естественно-языковых текстов можно с уверенностью сказать, что подавляющее большинство проблем лингвистической неоднозначности, которые представляют значительные, подчас непреодолимые сложности для современных систем автоматической обработки текстов, легко решаются естественным человеческим интеллектом на уровне подсознания. При этом никаких задач логической обработки или сложного перебора структур данных большой рекурсивной степени вложенности не происходит. Иначе подобная обработка происходила бы на уровне сознания и была бы явно осознанной деятельностью. Очевидно, что

решение основных проблем неоднозначности лингвистического анализа – полисемии, омонимии, онанимии – находится в другой плоскости. Решение этих задач напрямую связано с моделированием ассоциативно-семантической близости концептов в тексте, ассоциативно-смыслового контекста, влияющего на семантическую интерпретацию языковых знаков, моделированием ассоциативного веса концептов в семантической структуре текста и других явлений и процессов «правостороннего» типа, позволяющим описывать и алгоритмически реализовывать работу «подсознательного» ассоциативного связывания слов и концептов в лингвистическом анализе. Именно этим и мотивирована сегодняшняя чрезвычайная актуальность ассоциативно-семантических методов и алгоритмов в компьютерной лингвистике. В других направлениях искусственного интеллекта подобные задачи, не имеющие эффективных решений в рамках «левосторонних» логико-рекурсивных моделей, также успешно решаются аналогичными методами «правого» ассоциативно-семантического подхода.

Целью данной работы является разработка и анализ новых алгоритмов и методов автоматического реферирования естественно-языковых текстов. Стандартные методы частотного реферирования при обработке текста не могут учитывать сложную форму строения и особенности его синтаксических и семантических структур. И как следствие, классические алгоритмы определения частотных характеристик и зависимостей внутри текстового лексического корпуса, генерируя реферат посредством выборки отдельных предложений из входного текста, принципиально не в состоянии порождать целостный связный минимизированный текст реферата. Для решения проблемы автоматического построения связного реферата нужно задействовать новые принципы минимизации и оптимизации на уровне семантических структур входного текста, учитывая критерии связности и целостности естественно-языковых текстов. Для этого необходимо формализировать эти понятия.

Рассмотрим семантическую структуру входного текста, которая генерируется на этапе семантико-онтологического анализа и поступает на вход программного блока семантической постобработки текста. Семантическая структура входного текста является графом, вершины которого являются семантическими концептами текста (концептами действия, объектами, свойствами и т.д.), а дуги – семантико-синтаксическими отношениями между концептами в предложениях. Центральным семантико-синтаксическим концептом в предложениях, согласно традиции, считается предикатный концепт-действие (в синтаксисе – глагол), он является вершиной семантико-синтаксической конструкции. По аналогии с математикой, первичным в структуре семантической формулы есть предикат, аргументы являются подчиненными объектами, но от их значения зависит конечное значение предиката и всей формулы.

Рефератом текста называют уменьшенный по объему текст, который имеет тот же смысл и содержание. Хотя содержание может быть несколько урезанным, но реферат текста должен давать полное представление об оригинале. То есть из текста отбрасываются малосущественные (с точки зрения восприятия смысла) куски.

Реферирование представляет собой процедуру минимизации семантической структуры исходного текста. Под минимизацией нужно понимать уменьшение объема графа. Причем удалять можно только **семантически маловесомые** вершины графа. Возникает проблема определения **семантически весомых** и **маловесомых** вершин и

дуг. Существует гипотеза, что концепт считается тем весомее, чем чаще он упоминается в тексте. То есть частотный анализ может полностью определить наиболее важные концепты текста, взвесить частотный вес всех концептов, а также вес каждого предложения, суммировав вес всех концептов в предложении. Отбросив маловесомые предложения, можно получить реферат в первом приближении [2].

Кроме частотного веса концептов текста существует также ассоциативный вес. Примем, что концепт является тем более важным для семантики текста, чем больше связанной есть соответствующая ему вершина в семантическом графе входного текста. Для каждой вершины нужно вычислить количество соседей в семантическом графе текста. Опыт показывает, что при вычислении ассоциативного веса эффективным методом является также пересчет соседей на глубину 2 транзитивно через концепт-предикат. Таким образом связываются два концепта-аргумента, которые имеют совместный предикат в предложениях текста. Пример: «Кот гонялся за мышью»; рассмотрев предикат *гоняться* (*кот, мышь*), ассоциативно свяжем концепты-аргументы *мышь* и *кот*.

При минимизации семантического графа возникает проблема определения способа удаления из него маловесомых подграфов. В качестве критерия определения важнейших вершин можно принять ассоциативный, частотный или комбинированный вес. Если производить удаление вершин с малым весом и их дуги, почти всегда возникает проблема с целостностью и связностью текста. Под целостностью понимают неразрывность «линии» связного текста [3].

Согласно наипростейшему определению, предложения выражают законченную мысль. Формально можно перефразировать, что это предикат или суперпозиция предикатов с заполненными аргументными полями. Все поля предикатов должны быть заполнены. В свою очередь, в абзаце последовательно расположены предложения, тесно связанные по смыслу. По аналогии можно сказать, что предикаты соседних предложений должны быть связаны через тождественные значения аргументов. Или в соседних предложениях идет речь об одном и том же самом процессе (одинаковое имя предиката, концепта-предиката), или же их объединяет субъект, объект, место, действие и т.д. Таким образом, становится очевидным, что предикаты, которые соответствуют соседним предложениям, связаны тождественными значениями некоторых аргументов или одним именем предиката.

Если рассмотреть текст как совокупность абзацев, то связи между предикатами внутри абзаца являются более сильными, чем связи между предикатами из смежных абзацев, что выражается большим числом тождественных значений аргументов, чем в разных абзацах.

Тогда подграф, который выражает семантику отдельного абзаца, есть таким подграфом, внутри которого вершины являются более связанными друг с другом, чем с «внешними» вершинами из других абзацев. Подграф абзаца имеет вид «густого клубка» на фоне более редкой семантической сети текста. Точнее, семантическая структура текста представляет собою сеть густых подсетей абзацев, связанных между собой существенно слабее, чем вершины внутри этих подсетей.

Сильнее всех связаны подграфы соседних предложений внутри одного абзаца, далее идут подграфы несмежных предложений внутри одного абзаца, потом подграфы из соседних абзацев и т.д.

Исходя из этого, можно предположить, что связностью текста есть, с одной стороны, связность предикатов предложений текста по значениям аргументов, с другой стороны – это отсутствие изолированных компонентов в семантической сети текста.

Сохранение целостности можно сделать одним из критериев при выборе маловесомых подграфов текста в процедуре минимизации семантической структуры.

Рассмотрим саму процедуру минимизации семантической структуры текста.

Вход: семантическая сеть текста.

Выход: минимизированная семантическая сеть текста.

1. Сначала вычисляется ассоциативный вес каждой вершины графа. Для этого нужно обойти весь граф и определить количество соседей для каждой вершины.

2. Далее нужно обойти весь граф и удалить все вершины, что имеют вес, меньший, чем некоторый установленный пороговый уровень. Если вершины имеют тип «действие», то необходимо удалять их вместе с инцидентными ребрами. Если в результате удаления этих ребер образовались изолированные вершины, их также нужно удалить (предикат нужно удалять вместе с аргументами, то есть с глаголом удаляется все предложение). При этом необходимо проверять выполнение условия связности графа. Если в результате удаления вершины или подграфа в сети создаются изолированные компоненты, от удаления нужно отказаться.

Как указано выше, после ассоциативного взвешивания, нужно установить пороговый уровень ассоциативного веса вершин графа. Это зависит от желаемого уровня компрессии текста. Именно он определяет, сколько вершин и соответственно дуг нужно удалить из графа. Разделим множество всех вершин на подмножества вершин с одинаковым значением ассоциативного веса. Количество вершин в каждом подмножестве известно. Это легко определяется при ассоциативном взвешивании графа.

Чтобы достигнуть необходимого уровня компрессии, нужно начать суммировать количества вершин каждого из подмножеств, начиная с самого легкого по весу подмножества и далее по возрастанию, пока сумма не будет больше, чем то количество вершин, что нужно отбросить. Вес вершин текущего подмножества будет искомым пороговым уровнем.

Пусть граф содержит N вершин. Нужная компрессия 50 %. Это значит, что должны быть удалены [N/2] вершин с наименьшим весом.

Пусть в результате процедуры ассоциативного взвешивания графа образован массив subsetno, где subsetno [i] – количество вершин в подмножестве с ассоциативным весом i.

```
1  s: = 0;
2  While s<= [N/2] do
3  begin
4  i: = i + 1;
5  s: = s + subsetno [i]
6  end;
7  r: = i - 1;
```

Переменная r содержит значение порогового уровня ассоциативного веса. Если вершина имеет вес меньше r, ее нужно удалить.

Далее необходимо обойти весь граф, удаляя вершины, которые не достигли порогового уровня. Вершины удаляются вместе с инцидентными ребрами. Когда мы

удаляем вершину предикатного типа, создаются изолированные вершины, которые соответствуют аргументам данного предиката. Очевидно, эти вершины также нужно удалять. Благодаря этому эффекту уровень компрессии выходит больше, чем запланированный.

На выходе алгоритма генерируется минимизированная семантическая сеть текста, являющаяся образом будущего реферата. Применив к ней алгоритмы синтеза текстов, получим текст реферата. Из исходного текста выбираются те семантико-синтаксические сегменты предложений и абзацев, которые соответствуют вершинам и дугам, оставшимся в минимизированном графе. Таким образом генерируется текст реферата.

Моделирование таких интересных и сложных явлений и процессов, как ассоциативные связи в тексте, семантическая близость, анализ ассоциативного контекста, ассоциативный вес слов и концептов в структуре текста открывает широкие перспективы и дает средства для решения до сих пор принципиально нерешаемых задач не только в рамках компьютерной лингвистики, но, как показывает практика, практически во всех сферах и направлениях искусственного интеллекта.

Литература

1. Анисимов А.В. Информатика. Творчество. Рекурсия. – К.: Наук. думка, 1988.
2. Скороходько Э.Ф. Семантические сети и автоматическая обработка текста. – К.: Наук. думка, 1984.
3. Лукашевич Н.В., Добров Б.В. Автоматическое выявление лексической связности текста // Труды Казанской школы по компьютерной и когнитивной лингвистике «TEL-2001». – Казань, 2001.

А.В. Анісімов, О.О. Марченко

Асоціативне реферування природно-мовних текстів

У статті розглядається побудова асоціативно-семантичного алгоритму реферування текстів на природній мові. Асоціативний аналіз і моделювання асоціативно-семантичних зв'язків у тексті відкривають широкі перспективи вирішення фундаментальних проблем комп'ютерної лінгвістики.

A.V. Anisimov, A.A. Marchenko

Associative Natural Language Text Abstracting

The developing of associative-semantic algorithm for natural language text abstracting is described in this article. Associative analysis and modeling of associative-semantic text connections give perfect perspective for solution of major fundamental problems in computer linguistics and artificial intelligence.

Статья поступила в редакцию 20.06.2006.