Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза

А.Е. Ермаков, к.т.н., ведущий разработчик ООО "Гаран-Парк-Интернет"

Аннотация

В докладе предлагается метод построения информационного портрета документа на основе элементов смысла текста, извлекаемых средствами синтаксического анализа и синтеза. Метод основан на использовании синтаксического анализатора с последующим преобразованием семантической сети во множество строк, которые представляют в унифицированном виде все элементарные отношения между сущностями в тексте и являются корректными с точки зрения грамматики русского языка. Описывается ряд преобразований синтаксических структур, обеспечивающих инвариантность представления смыслов от ряда особенностей поверхностно-синтаксической организации текста. Предлагается способ ранжирования полученных элементов смысла по информативности с точки зрения характеристики текста в прикладных системах.

Введение

При компьютерной обработке текста в информационно-поисковых системах возникает задача автоматического построения содержательного портрета документа, описывающего в унифицированной и компактной форме его основные смысловые атрибуты – фигурирующие в тексте ситуации и сущности, в них вовлеченные. Только наличие развитого смыслового портрета позволяет эффективно решать задачи аналитической обработки документа, такие как автоматическая классификация и интеллектуальный поиск информации. Для решения подобных задач уже существует множество алгоритмов, приложения которых описаны, в частности, в работах [1-3], однако их эффективность напрямую зависит от качества извлечения смысловых атрибутов текста – качества лингво-аналитического программного обеспечения.

Наиболее полным описанием смысла текста является семантическая сеть, формируемая на основе синтактико-семантического анализа. Семантическая сеть позволяет абстрагироваться от малоинформативных элементов формально-синтакисической структуры текста (порядка слов, залога и т.п.) и представляет его пропозициональную структуру в терминах описываемых ситуаций (предикатов) и их участников (аргументов) в определенных семантических ролях.

Однако полное представление смысла текста в форме семантической сети является избыточным и непродуктивным для многих практических задач, имеющих ограничения на вычислительные ресурсы. Такое представление имеет большой объем (превышающий объем документа), а его утилизация требует развитых нетривиальных средств для поиска и сравнения структур на графах.

Более экономичным и удобным является описание смыслового портрета в форме перечня элементарных смыслов - атрибутов, с оценками их информативности для характеристики текста. Традиционно в силу простоты реализации для этой цели используются частотные списки употребляющихся в тексте слов, однако очевидно, что наиболее информативные элементы смысла, описывающие отношения, возникают только на уровне синтагм, выделение которых требует применения нетривиальных алгоритмов синтаксического анализа, описанных, например, в [4]. Будучи дополнен правилами для генерации канонической формы синтагм, синтаксический анализ-синтез позволит описать каждый смысловой атрибут текста в виде строки, инвариантной к его грамматическому выражению в различных фразах.

Например, фразам "*Транспорт был арендован предприятием у автобазы*", "*Предприятие арендует у автобазы транспорт*" и "*Аренда транспорта предприятием у автобазы*" будут соответствовать одинаковые элементы смысла: "*предприятие арендует*", "*аренда транспорта*", "*аренда у автобазы*".

1. Синтаксический анализ и эксплицирование отношений

Синтаксические связи между словами в тексте можно разделить на три общих класса:

1. связи между ситуациями и их участниками – предикатно-аргументные связи (*подписать -> указ*, *покупка -> земли*, *договор -> (о) разоружении*). Описываются моделями управления предикатов.
2. связи внутри именных групп, обычно называющих участников ситуации – атрибутивные связи (*новый <- указ -> президента, человек -> (с) ружьем*). Описываются общими правилами грамматики языка.
3. связи между ситуациями - предикатно-предикатные (*учиться -> читать*, *видеть -> (как) обнаружили, идти -> качаясь*). В основном описываются моделями управления предикатов второго порядка.
4. связи ситуаций с обстоятельствами или дополнительными атрибутами - сирконстантные связи. Выражаются предложно-падежными формами существительных либо наречиями и эквивалентными им формами прилагательных.

В первую очередь для характеристики смысла текста значимы связи первых двух классов.

Для выделения всех связей необходимо использование синтаксического анализатора русского языка, который подбирает оптимальное покрытие синтаксической структуры фразы набором правил, описывающих элементарные синтагматические отношения между словами текста [4]. Получаемое в результате анализа дерево синтаксических зависимостей должно быть преобразовано в семантическую сеть, содержащую все присутствующие попарные семантические отношения между словами. Для этой цели необходимо использование дополнительного комплекса правил, позволяющего эксплицировать все отношения, скрытые в дереве синтаксических зависимостей.

Например, протагониста действия, выраженного глаголом, следует дублировать на синтаксически зависимое деепричастие или подчиненные члены глагольной группы: "*начальник <- подписал -> сомневаясь*" = "*начальник <- подписал*, *начальник <- сомневается, подписал -> сомневаясь* ".

Наиболее часто эксплицирумым видом отношений являются отношения с однородными членами именных и глагольных групп: "*честь и достоинство <- опозорить и утратить*" = "*честь <- опозорить*, *достоинство <- опозорить*, *честь<- утратить*, *достоинство <- утратить*".

На этом же этапе необходима замена местоимений (в том числе анафорических) их референтами: *"... Вася, который <- обидел -> себя*" = "*Вася <- обидел, обидел -> Васю*".

Следует отметить, что и после преобразований все связи семантической сети являются направленными и отражают те синтаксические зависимости, которые могли бы явно присутствовать в эквивалентном по смыслу тексте, т.е. глагол подчиняется только глаголу, существительное – глаголу или существительному, а прилагательное – существительному или глаголу. Здесь мы принимаем, что протагонист (даже в роли подлежащего) всегда подчиняется предикату, однако допустимо принять и противоположное направление связи.

2. Синтаксический синтез элементов смысла

Для формирования упрощенного смыслового портрета текста, как указывалось выше, следует провести синтез строк, представляющих выделенные элементарные отношения в некоторой канонической форме, которая позволит их однозначно отождествлять и будет естественной для восприятия человеком.

В силу свойств графа семантической сети любой путь на нем в направлении связей от главного слова к зависимому представляет некоторый законченный и самостоятельный смысл, который может быть развернут в линейную, синтаксически правильную структуру. С этой целью можно применить ряд правил синтеза, каждое из которых способно сформировать каноническую форму для пары связанных слов. В русском языке грамматическая форма любого слова полностью определяется типом отношения с тем единственным словом, от которого оно зависит (в тексте и семантической сети). Таким образом, процесс синтеза цепочек связанных слов любой длины в направлении от главного слова к зависимому полностью однозначно описывается отдельными правилами синтеза форм для каждого из отношений. Применяя их рекурсивно и проходя все цепочки связей от каждого из слов, можно получить конечное множество строк, выражающих все элементы смысла - синтагм.

Для синтеза корректных с точки зрения русского языка синтагм, эксплицирующих отношения классов (1) и (2), достаточно выполнять следующие общие правила:

- падеж первого главного слова, с которого начинается синтез цепочки – именительный (для существительных).
- каждое правило задает позицию зависимого слова и его падеж. Зависимое прилагательное ставится слева от главного слова, прочие слова – всегда справа. Падеж зависимого слова (и предлог, если есть) для предикатно-аргументных (или сирконстантных) связей определяется моделью управления предиката, в соответствии с которой была установлена связь при анализе. Для атрибутивных связей с подчиненными прилагательными наследуется падеж существительного, с существительными без предлога – родительный, а для существительных с предлогом падеж определяется предлогом, как в тексте.
- все глаголы ставятся в форму инфинитива за исключением сочетаний с существительным в роли протагониста, где они ставятся в личную форму, согласованную по роду с существительным (настоящего, а если нет – прошедшего времени).
- все слова ставятся в форму единственного числа (если таковой нет - множественного), а зависимые прилагательные согласуются по роду и падежу с существительным.

Отметим, что отождествление конструкций с пассивным и активным залогом осуществляется ранее на этапе преобразований синтаксической структуры в семантическую сеть, когда именительный в пассиве переходит в винительный в активе, а творительный – в именительный: “*указ(И) подписан президентом(Т)*” = “*президент(И) подписал указ(В)*”. Аналогично производится преобразование падежей в причастных оборотах.

Таким образом, синтез ведется с опорой не “синтаксические” падежи, употребленные в тексте, а на семантические, которые определяются моделями управления.

Применение тезауруса, содержащего синонимы как из одной, так и из разных частей речи, позволяет привести глагольные формы предикатов к эквивалентным существительным (“*купить -> книжку*” = “*покупка книги*”). При этом для определения падежа зависимого существительного используются следующие правила:

- падеж прямого дополнения - родительный ("*критикует журналиста*" = "*критика журналиста*").
- падеж логического подлежащего при переходном глаголе – творительный ("*критикует журналист*" = "*критика журналистом*", при непереходном – родительный ("*летит самолет*" = "*полет самолета*").
- В прочих случаях предложного и беспредложного управления падеж зависимого слова сохраняется.

Например, в результате обработки фразы: "*Арбитражный суд в этом году будет рассматривать договор об аренде земли, подписанный директором металлургического комбината и муниципалитетом*" могут быть выделены следующие смыслы, выраженные словосочетаниями:

От предиката "*рассмотреть*" -
"*Рассмотрение договора об аренде земли*", "*Рассмотрение договора об аренде*", "*Рассмотрение договора", "Рассмотрение арбитражным судом*", "*Рассмотрение судом*"

От предиката "*подписать*" -
"*Подписание договора об аренде земли*", "*Подписание договора об аренде* ", "*Подписание договора*", "*Подписание директором металлургического комбината*", *"Подписание директором", "Подписание директором комбината*", "*Подписание муниципалитетом*"

От предиката "*договор*" -
" *Договор об аренде земли*", "*Договор об аренде"*

От предиката "*аренда*" -
"*Аренда земли*"

От существительного "*директор*" -
"*Директор металлургического комбината*", "*директор комбината*"

От существительного "*комбинат*" –
"*Металлургический комбинат*"

От существительного "*суд*" –
"*Арбитражный суд*"

3. Ранжирование элементов смысла и информационный портрет текста

Как показано выше, обход графа семантической сети позволяет эксплицировать основные элементы смысла текста, и сформировать детальный информационный портрет документа. Однако очевидно, что не все элементы обладают одинаковой информативностью с точки зрения характеристики текста, и в ряде прикладных задач, требующих сравнения документов по смыслу, требуется учесть это обстоятельство.

При оценке информативности предлагается учитывать два факта:

- объективную структурную организацию (иерархию) смыслов, которая выражается в уровне синтаксической зависимости одних элементов от других. Например, смыслы, входящие в состав синтагмы "*встреча президента России*", не равнозначны: речь идет в первую очередь о "*встрече*", затем о "*президенте*", и лишь опосредованно затрагивает "*Россию*". В тоже время цельный смысл "*встреча президента России*" более информативен, чем "*встреча президента*" и просто "*встреча*", так как включает в себя конкретизирующие элементы.
- субъективную коммуникативную организацию смыслов, которая выражается в различии синтаксических ролей элементов во фразе, назначаемых автором при порождении текста. Например, в конструкциях типа "*состоялась встреча президента ...*" и "*на встрече президента состоялось ...*" рассматриваемый смысл имеет различную значимость для автора сообщения: в первом случае он соответствует теме (центру) события и выражается синтаксической ролью

подлежащего, а во втором представляет обстоятельство, на фоне которого разворачивается иное, более значимое событие.

Простейшую количественную оценку информативности может дать следующая формула:

$\omega = Wi\ Wi^*\ Li / Li^*$, где

Li - длина пути на графе семантической сети, порождающей i-й смысл,
Wi – вес синтаксической роли главного слова в пути, порождающем i-й смысл,
а Li* и Wi* - длина и вес синтаксической роли соответственно для наиболее длинного пути, включающего путь, порождающий i-ый смысл.

Классификация синтаксических ролей может включать следующие типы: подлежащее, сказуемое, прямое дополнение, второстепенное сказуемое, косвенное дополнение, определение (в составе именных групп), обстоятельство. В целом указанный порядок хорошо отражает коммуникативную иерархию элементов смысла для автора текста (хотя бывают исключения). Выбор конкретных значений весов определяется эмпирическими соображениями и может зависеть от особенностей решаемой прикладной задачи.

Заключение

Описанные методы анализа-синтеза элементов смысла текста и методика их ранжирования разработаны в компании “Гарант-Парк-Интернет” и внедрены в ряд коммерческих продуктов, выпускаемых под торговой маркой RCO. С подробной информацией можно познакомиться на сайте *http://www.rco.ru*.

Литература

1. Ермаков А.Е. Проблемы полнотекстового поиска и их решение. // Мир ПК. – 2001. – N 5. – С. 64-66.
2. Ермаков А.Е., Плешко В.В. Тематическая навигация в полнотекстовых базах данных. // Мир ПК. – 2001. – N 8. – С. 52-55.
3. Плешко В.В., Ермаков А.Е., Липинский Г.В. TopSOM: визуализация информационных массивов с применением самоорганизующихся тематических карт // Информационные технологии. - 2001. - N 8. – С. 8-11.
4. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста. // Информационные технологии. - 2002. – N 7. – С. 30-34.