Синтаксический анализатор русского языка на основе данных морфологии: функциональный подход

Евгений Черкашин[1,3], Наталия Свердлова[2] and Елена Марьясова[2]

[1]*Институт динамики систем и теории управления СО РАН, Ул. Лермонтова, д. 134, Иркутск, 664033, Россия*
[2]*Иркутский научный центр СО РАН, Ул. Лермонтова, д. 134, Иркутск, 664033, Россия*
[3]*Институт математики и информационных технологий Иркутского государственного университета, Бульв. Гадарина, д. 20, Иркутск, 6640003, Россия*

Аннотация
Задача исследования состоит в создании технологии синтаксического анализа предложений русского языка при помощи обработки морфологических данных слов из корпуса языка АОТ, функций-предикатов распознавания соответствия слов по ггамматическим (синтаксическим) правилам, управляемого эвристическим переборным алгоритмом, реализованным на компилируемом функциональном языке программирования Haskell.

1. Введение

Задача распознавания синтаксической структуры (РСС) русского и английского текста в документах - одна из часто встречающихся. Требования к качеству РСС предложений варьируется от задачи к задаче. В одних документах требуется объединить строки, формирующие абзац, в предложения, а в других - определить характеристики предложений, а, иногда, и смысл изложенного. Если не требуется полного соответствия грамматике языка, то достаточно применить алгоритм эвристического определения конца предложения, выдающего правильный ответ с некоторой высокой вероятностью 0.95 [1]. Более глубокий вид анализа представляет собой технология link-grammar (грамматика связей) [2], подход достаточно универсальный, и, в целом, независим от структуры языка. Идея link-grammar состоит в представлении модели связей каждого слова в предложении с другим словом при помощи вариантов сцепления. Авторы работы [2] показали, что основным ограничением, наиболее вероятно, имеющим естественные причины возникновения, является тот факт, что получаемый граф связей должен быть планарным. Наборы слов, для которых невозможно построить такой граф не воспринимаются человеком как предложение, несущее смысл. Еще одним направлением РСС является построение дерева синтаксического разбора (ДСР), состоящего из направленных связей)

6^{th} International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2024), July 01–05, 2024, Irkutsk, Russia
eugeneai@irnok.net (Черкашин); nsverdlova@yandex.ru (Свердлова); mariaselena@yandex.ru (Марьясова)
0000-0003-2428-2471 (Черкашин); 0000-0002-5315-6266 (Свердлова); 0000-0002-3504-9416 (Марьясова)

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)

на основе грамматических зависимостей (dependency grammar).

Целью данного исследования - реализация подхода к РСС использующего грамматические зависимости, где связи между словами формируются по правилам русского языка. Связь считается возможной, если выполняются ограничения на морфологическое соответствие слов, входящих в связь (правило), и полученный граф связей предложения также является планарным. В отличие от подходов [3] варианты связей в предложении строятся только на основе данных корпуса русского языка, в виде варианта построения нагруженного И-ИЛИ-дерева вариантов распознавания. При внесении дополнительных модификаций данный алгоритм может быть расширен на решение задачи интерпретации дейктических связей между предложениями, например, указание местоимением существительного.

2. Методика представления процесса распознавания синтаксической структуры предложения

Первым этапом проекта является синтаксический анализ предложения. Подход, реализуемый на данном этапе, основан на восходящем распознавании синтаксической структуры по правилам русского языка. В результате строится планарное дерево (набор вариантов таких деревьев) синтаксического разбора, ассоциированное с оценкой вероятности реализации данного разбора. В качестве языка реализации выбран Haskell, позволяющий задавать правила грамматики виде функций-предикатов на основе сопоставления с шаблоном, входной поток так называемых граммем - в виде потенциально бесконечного списка, процесс трансляции - преобразование списка граммем в список синтаксических структур, варианты разбора - списком процессов, и, самое главное, за счет ленивой модели вычисления производить обработки этих потенциально бесконечных структур. Кроме того, Haskell порождает скомпилированный запускаемый модуль. Исходный поток граммем получается из внешнего локального сервиса, реализованного на языке Python и библиотеки Pymorphy2 [4].

Приведем пример задания правил распознавания, представленных в виде функций join, join3 класса Rule.

```
instance Rule GRAM where
  join :: GRAM -> (Gram -> Gram -> Bool,
                   Gram -> Bool, Gram -> Bool)
  join AdjNoun = (adjNounConsist, isAdj, isNoun)
  join NumrNoun = (numrNounConsist, isNumr, isNoun)
  join SubjVerb = (subjVerbConsist, isSubj, isVerb)
  join VerbTranObjAccs = (isAnyRel, isVerbTran, isObjAccs) -- (1)
  join NounNounGent = (isAnyRel, isNoun, isNounGent)
  join Percent = (isAnyRel, isNum100, isPercent)
  join PhoneNumber = (isAnyRel, isWord "+", isPhoneNumber)
  join Sentence = (isAnyRel, hasWall, isSentenceEnd)
  join AdvbVerb = (isAnyRel, isVerb, isAdvb)
```

```
-- . . . . . .
  join _ = (lfm, lf, lf)
    where lf _ = False
          lfm _ _ = False

  join3 :: (GRAM, String) ->
           (Gram -> Gram -> Bool, Gram -> Bool, Gram -> Bool)
  join3 (ForJoin, "для") = (isAnyRel, isNoun, isNounGent)
  join3 (NounInNoun, "в") = (isAnyRel, isNoun, isNounLoct)
-- . . . . . .
  join3 _ = (lfm, lf, lf)
    where lf _ = False
          lfm _ _ = False
```

Структура GRAM представляет собой множество грамматических характеристик слов корпуса русского языка АОТ (Автоматическая обработка текста) [5], а также обозначения грамматических связок, все конструкторы являются константами. Функция join ассоциирует конструктор GRAM с тремя предикатами - а) двухаргументный предикат, задающий ограничение на совместное наличие характеристик “склеиваемых” в грамматическую структуру слов, б) два одноаргументных предиката, задающих индивидуальные, независимые характеристики склеиваемых слов. Предикат isAnyRel обозначает, что на слова не задаются совместные ограничения, т.е. для любых входных слов возвращает True.

В тексте примера меткой (1) выделено правило распознавания переходного глагола с существительным в винительном падеже. Здесь не накладывается совместных ограничений. Рассмотрим пример разбора простейшего предложения “мама мыла раму”. Граммемы pymorphy2 слов следующие (вывод сокращен):

мама:
1. word='мама', 'NOUN,anim,femn sing,nomn', nf='мама', score=1.0,

мыла:
1. word='мыла', 'NOUN,inan,neut sing,gent', nf='мыло', score=0.333
2. word='мыла', 'VERB,impf,tran femn,sing,past,indc', nf='мыть', score=0.333
3. word='мыла', 'NOUN,inan,neut plur,nomn', nf='мыло', score=0.166
4. word='мыла', 'NOUN,inan,neut plur,accs', nf='мыло', score=0.166

раму:
1. word='раму', 'NOUN,inan,femn sing,accs', nf='рама', score=0.888
2. word='раму', 'NOUN,inan,masc,Geox sing,datv', nf='рам', score=0.11

Разбор начинается с первой пары слов предложения. Первые два слова, которые можно объединить - это “мама” и “мыла”. Граммемы слов порождают четыре комбинации, которые фильтруются предикатами правил: производится неэвристический перебор правил, запускаются функции-предикаты. Правила, предикаты которых истинные, задают вариант связи. Оценка связи делается перемножением значения score участвующих в связи слов. В нашем примере только одно правило остается - “подлежащее-глагол” (SubjVerb), где подлежащее должно

быть существительным (NOUN, femn, sing) в именительном падеже (nomn), глагол (VERB) соответствовать роду и числу подлежащего (femn, sing). Формируется связь с оценкой 0.333. Следующее слово - “раму”, оно может быть соединено или с “мыла” или с “мама”. Правило согласования “мыла раму” - переходный (tran) глагол (VERB) с существительным (NOUN) в винительном падеже (accs) (VerbTransObjAccs). В результате получается две последовательные связи в предложении, состоящем из трех слов. Общая оценка разбора - сумма оценок всех сформированных связей, достаточная чтобы можно было сравнить варианты разбора.

Правила, заданные в программе, предполагают прямой порядок слов в предложении (SVO). Если для очередного слова не удается построить связь с предыдущим словом (обратный порядок слов, вводная фраза, отсутствие правила) порождается три альтернативных варианта: а) создание связи общего вида - “слабый” вариант разбора, б) пропуск слова и формирование процесса РСС вводной фразы, в) помещение слова в список слов, расположенных в обратном порядке для последующего продолжения анализа по основному варианту. Оценка связи для варианта (а) равняется 0.1 и уменьшается в 0.7 раза для всех последующих таких связей. В случае (б) строится отдельная фраза.

3. Развитие средств анализа текста

Основная цель проекта – разработать методику машинного перевода научного текста с одного естественного языка на другой, привязанного к заданной предметной области. Одним из требований выступает необходимость реализации системы логического обоснования принятых решений. Процесс перевода включает несколько этапов: 1) анализ синтаксической структуры предложения, 2) выделение в предложении отдельных фраз и связывание их в древовидные синтаксические структуры общего вида (ДСОВ), 3) привязка фраз к концептуальной модели предметной области текста, 4) интерпретация ДСОВ исходного языка в ДСОВ целевого, 5) порождение предложения целевого языка. В качестве исходного языка выбран русский язык, а целевого - английский (и наоборот), как языки представляющие наибольший интерес в части преподавания профессионального перевода.

Работы ученых в области РСС в 80х-90х годах 20 века показали, что качественный перевод, реализуемый без использования машинного обучения, требует, чтобы программа не только анализировала синтаксическую структуру предложения, отображала термины при помощи словаря, но и понимала в некоторой степени суть описанного, потенциально, задавала вопросы с целью уменьшения неопределенности, высказывала несогласие с изложенным. Стандартный подход к моделированию некоторой знаковой системы, включая автоматический перевод, включает задание трех моделей: семантической, синтаксической и модель прагматики. Соответственно, для реализации качественного перевода необходимо формировать модель предметной области до начала реализации перевода, т.е. семантику и прагматику, а также алгоритмы привязки синтаксических структур

к этим моделям.

Модели семантики и прагматики переводчика предполагается задавать при помощи расслоения [6] онтологий, концептуальных моделей, представленных явно соответствующими средствами при помощи формального языка. Каждый слой расслоения представляет некоторый аспект предметной области, между слоями задаются морфизмы, отображающие элементы одной онтологии, более специализированной, на более общую. Слои задаются средствами моделирования онтологий (например, Protege, UML, mind maps), а морфизмы - при помощи правил Prolog и отношений OWL2. Кроме того, слой и совокупность слоев представляется в виде объекта Logtalk, объектно-ориентированной надстройки над языком Prolog. Использование Logtalk позволяет инкапсулировать модели семантики и прагматики в рамках одного объекта, интерпретируемого, в частности, как фрейм Марвина Мински.

4. Варианты использования в обучении профессиональному переводу

Обучение профессиональному переводу научных текстов связано с рядом проблем, частичное решение которых возможно при помощи машинного перевода. Например, в задаче обучения русскоговорящих студентов иностранному языку научного общения, так и иностранных студентов русскому языку, необходимо формирование базиса научной коммуникации. Программы на основе технологий синтаксического анализа, будучи использованными в учебном процессе, поспособствуют пониманию структуры предложения и его грамматики, в том числе типы предложений, порядок слов, основные синтаксические конструкции и правила их построения, связи между лексическими единицами. В результате повышается эффективность процесса обучения иностранному языку, ускоряется приобретение навыков научного общения, позволяя студентам лучше понимать и овладевать структурой и особенностями изучаемого языка.

Автоматический анализ больших объемов текстов позволяет повышать качество перевода, выявляя языковые закономерности, тенденции и особенности контекста. Данные такого анализа могут использоваться для лингвистических исследований, анализа структуры научных текстов и изучения различий между разными стилями общения. Одно из прикладных направлений - это автоматическое создание рефератов, аннотаций или инструменты для коррекции структуры и грамматики текстов.

На базе разрабатываемых технологий аспиранты и молодые ученые могут создавать собственные языковые продукты, в частности научные статьи, как на русском, так и на иностранном языках, при автоматизированной поддержке обнаружения грамматических ошибок, структурных несоответствий в текстах и не коррелирующих с исходной интенцией конструкций, что, в сою очередь, повышает языковую грамотность.

При работе над научными статьями на иностранном языке синтаксический

анализатор будет эффективным инструментом перевода и адаптации своих исследований. Это позволяет более точно передать на иностранном языке смысл и структуру предложений из русскоязычного оригинального текста, что важно для сохранения точности и качества перевода. При этом обеспечивается грамматическая корректность и логическая последовательность предложений, оценивается и поддерживается стилистическое соответствие текста. Разрабатываемые инструменты могут значительно облегчить процесс создания научной статьи на иностранном языке, повышая профессионализм и качество исследовательских работ молодых ученых и аспирантов.

Исследования в области лингвистики также могут быть поддержаны инструментально, например, в задаче выделения ключевых терминов и специализированной терминологии в научных текстах определенной научной специальности. Здесь важно создать актуальный глоссарий научного дискурса, включающий все релевантные термины и определения для конкретной предметной области. РСС позволяет определить синтаксические связи и отношения между различными терминами в тексте, сформировать полные и точные определения и связи между терминами в глоссарии. В конечном итоге указанные технологии обеспечивают точность, структурирование и организацию информации о терминах и их отношениях, облегчая понимание и обмен информацией в рамках научного институционального дискурса.

Формирование предметной области в сознании носителя научной информации как компонента системы адаптивного обучения является ведущим условием самостоятельного моделирования научно-исследовательской деятельности. Таким образом, полученный набор имманентных корпусу научных текстов свойств будет использован для создания нового языкового продукта, который представляет собой языковой инвариант - готовую семантико-синтаксическую основу для описания процесса и результатов исследования в лингвистике.

5. Заключение

В статье представлены результаты начального этапа разработки предположительно однопроходного синтаксического анализатора предложений и фраз русского языка, адаптирующего базовые принципы алгоритма link-grammar [2] и грамматик зависимостей к вычислительной модели функционального языка программирования Haskell. Входной поток слов предложения (фразы), охарактеризованных так называемыми граммемами, представляется как бесконечный список структур. Данные структуры преобразуются в список грамматических связей, одновременно формирующих дерево грамматического разбора. Варианты построения синтаксического разбора оцениваются. Оценка позволяет задавать предпочтения того или иного варианта, устанавливать приоритет процесса дальнейшего разбора.

На данном этапе проекта производится уточнение структуры алгоритма, анализ литературы по грамматике русского языка, формализация правил языка, реализация библиотеки-аналога pymorphy2 [4] средствами Haskell с ее адаптацией

к режиму ленивых вычислений. Реализация технологии направлена на решение задачи распознавания текста в неразмеченных документах формата PDF или результатов сканирования/распознавания изображений.

Следующие этапы исследования связаны с построением моделей семантики и прагматики в виде полиситемного расслоения [6], частично представленного в проекте OpenCYC [7], а также правил распознавания отношений в дереве синтаксического разбора, реализованных, например в [8].

6. Благодарности

Исследование проведено при поддержке Министерства науки и высшего образования Российской Федерации, проект ИНЦ СО РАН «Лингвосемиотическая гетерогенность научной картины мира: теоретическое и лингводидактическое описание» (FWSN-2022-0001), № госрегистрации, и ИДСТУ СО РАН “Методы и технологии облачной сервис-ориентированной цифровой платформы сбора, хранения и обработки больших объёмов разноформатных междисциплинарных данных и знаний, основанные на применении искусственного интеллекта, модельно-управляемого подхода и машинного обучения”, (FWEW-2021-0005).

Список литературы

[1] Sentence boundary disambiguation. “Doing Things with Words, Part Two: Sentence Boundary Detection”, URL:https://en.wikipedia.org/wiki/Sentence_boundary_disambiguation#cite_note-2 (access date: 26.05.2024)

[2] D. Grinberg, J. Lafferty, D. Sleator. A robust parsing algorithm for link grammars. Carnegie Mellon University Computer Science technical report CMU-CS-95-125, and Proceedings of the Fourth International Workshop on Parsing Technologies, Prague, September, 1995. URL:https://www.cs.cmu.edu/afs/cs.cmu.edu/project/link/pub/www/papers/ps/tr95-125.pdf (access date: 26.05.2024)

[3] Ru-eval: оценка методов автоматического анализа текстов. URL:https://ru-eval.github.io/syntax.html (access date: 26.05.2024)

[4] Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, 2015, pp 320-332.

[5] Автоматическая Обработка Текста. URL:http://aot.ru/ (access date: 26.05.2024)

[6] А. Черкашин, Полисистемный анализ и синтез : Прил. в географии / А. К. Черкашин; Отв. ред. В. С. Михеев;. - Новосибирск : Наука : Сиб. предприятие, 1997. - 499 с

[7] The OpenCyc Platform. URL:https://github.com/asanchez75/opencyc (access date: 26.05.2024)

[8] J Apresian, I Boguslavsky, L Iomdin, et al. ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT. First International Conference on Meaning–Text Theory (MTT’2003), pp. 279-288