

Модуль фрагментарного анализа в составе системы машинного перевода Crosslator 2.0

Жирнов Р. В.

Институт Прикладной Математики им. М.В. Келдыша РАН

Россия, 125047, Москва, Миусская пл., д.4

ultro@mail.ru

Клышинский Э. С.

Институт Прикладной Математики им. М.В. Келдыша РАН

Россия, 125047, Москва, Миусская пл., д.4

klyshinsky@mail.ru

Максимов В. Ю.

Институт Прикладной Математики им. М.В. Келдыша РАН

Россия, 125047, Москва, Миусская пл., д.4

vadimmax2000@mail.ru

В данной статье описан модуль фрагментарного анализа, работающий в составе системы машинного перевода Crosslator 2.0. С помощью этого модуля в предложении выделяются фрагменты определенного типа. После идентификации фрагментов в предложении частично снимается омонимия и расставляются метки, что облегчает дальнейший синтаксический анализ. При этом в несколько раз возрастает скорость и улучшается качество перевода. Модуль действует с помощью коллекции правил, которые составляются для исходного языка.

Введение

В системах машинного перевода синтаксический анализ сталкивается с серьезнейшей проблемой генерации огромного числа вариантов разбора, особенно при разборе длинных предложений. Проблема возникает из-за наличия омонимии в естественном языке и неопределенностью при заполнении грамматически допустимых валентностей.

С учетом того, что с добавлением омонимов в предложении число вариантов разбора растет как декартово произведение, полный разбор длинных предложений одними только средствами синтаксического анализа зачастую вообще представляет

собой практически невыполнимую задачу. Резко увеличивается объем оперативной памяти и время разбора предложения, что, естественно, неприемлемо для пользователей. Конечно, при анализе можно уменьшать глубину разбора предложения, но это отрицательно сказывается на качестве разбора предложения.

Некоторые фрагменты предложения позволяют устранить исходную неоднозначность, если учитывать их структуру и место в предложении. Выделение таких фрагментов, частично снимающих омонимию, позволяет существенно сократить число рассматриваемых вариантов. При этом время разбора предложения может снижаться более чем на порядок.

Введение модуля фрагментарного анализа в систему машинного перевода между стадиями графематического и синтаксического анализа значительно повышает качество и сокращает время перевода. Если предложение удастся разбить на фрагменты, каждый из которых успешно проходит все стадии грамматического анализа, то мы получаем возможность разбирать достаточно длинные предложения за приемлемое время.

Актуальность данной проблемы подтверждается параллельными работами [4].

Определение фрагмента

Старший грамматический элемент естественного языка – это предложение, которое может состоять из одного или нескольких простых предложений или оборотов. Простые предложения и обороты в свою очередь состоят из одной или нескольких фраз, которые состоят из одного или нескольких слов, слова же состоят из одной или нескольких морфем (где морфема – минимальный грамматический элемент) [3].

Фрагментом же, с лингвистической точки зрения, считается одно или несколько слов, порядок и форма которых грамматически задана. Т.е. фрагментом может быть как одно слово, так и фраза, оборот или простое предложение в целом. В данной реализации фрагментарного анализа было использовано именно это определение фрагмента, с тем лишь уточнением, что в нашей системе фрагментом считается лишь такая последовательность слов, к которой применимо какое-либо

правило или правила из коллекции для данного естественного языка. Более подробно о правилах и их коллекциях будет рассказано ниже.

Задача фрагментарного анализа – устранять неоднозначности и передавать на синтаксический анализ однозначно определенные фрагменты текста, которые разбираются далее в синтаксический анализ единственно возможным способом. Это позволяет упростить синтаксический разбор с точки зрения быстродействия, расходов оперативной памяти и т.п. и улучшить качество перевода.

Правило как основной элемент фрагментарного анализа

Как уже было сказано выше, основой работы данной реализации фрагментарного анализа являются правила, служащие для выделения фрагментов во входном предложении. Каждое такое правило создается эмпирически на основе знаний о грамматике данного естественного языка и с учетом особенностей работы грамматического анализа системы Crosslator 2.0, в составе которой и функционирует модуль фрагментарного анализа.

Прежде чем приступить к описанию структуры правила, стоит отметить, что на вход фрагментарного анализа попадает предложение, где каждое слово представляет собой омоним, в каждом элементе которого уже определена часть речи и грамматические параметры (такие как форма, род и т.п.). Метки попадают во входное предложение в процессе работы фрагментарного анализа, т.е. некоторые из его правил могут маркировать определенные фрагменты метками, которые в свою очередь могут быть использованы другими правилами фрагментарного анализа, либо используются на дальнейших этапах работы системы машинного перевода.

Каждое правило фрагментарного анализа имеет следующую структуру:

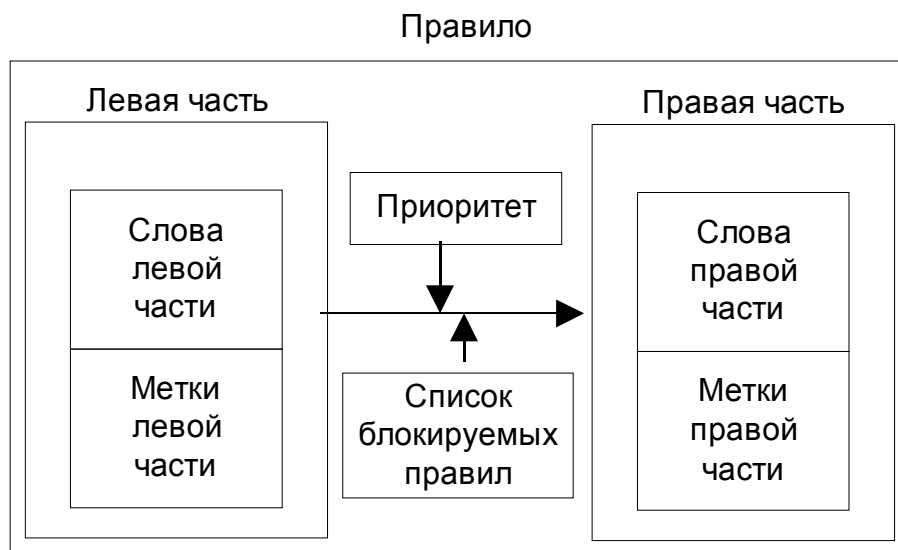


Рис. 1. Структура правила фрагментарного анализа.

Как видно из рисунка 1 правило имеет левую и правую части, а также приоритет и список блокируемых правил. Опишем все составляющие части правила подробнее.

Левая часть правила содержит слова и метки, которые и определяют применимость данного правила к фрагменту входного предложения. Слова в левой части сохраняются с учетом их порядка, равно как и метки. Причем существует возможность определения фрагментов по меткам и словам одновременно. Например, фрагмент может начинаться с метки 1, после неё должен содержать последовательность “слово 1, слово 2” и заканчиваться меткой 2. Для слов в левой части вместо полного описания слова (инфинитив, часть речи и параметры) есть возможность указать только часть речи (например, подходят все существительные), либо указать обратные признаки (подходят все слова кроме глаголов) и указать требуемые грамматические параметры (найти все глаголы прошедшего времени). Таким образом, если левая часть правила полностью совпала с неким фрагментом входного предложения, то данное правило считается применимым к фрагменту.

Правая часть правила содержит слова и метки, которые изменяются либо добавляются во входное предложение после определения того, как левая часть данного правила совпала с неким фрагментом. В словах правой части указывается, какие слова из фрагмента на что заменить, а какие полностью скопировать.

Иными словами, если некий фрагмент исходного текста подходит под описание левой части правила, в этом правиле мы можем расставить метки на

словах фрагмента, пронумерованных относительно начала или конца фрагмента. Например, в правой части правила мы можем поставить метку, начинающуюся на том слове или знаке препинания, который непосредственно предшествует началу фрагмента, и заканчивающуюся на последнем слове этого фрагмента. Таким образом, метка может устанавливаться на одно или несколько слов исходного предложения.

Предусмотрены даже такие варианты, в которых начало или конец метки могут быть не определены. В таком случае на синтаксический анализ приходит информация о том, что на данном элементе предложения фрагмент начинается или, наоборот, заканчивается.

Приоритет правила – это число, которое показывает преимущество данного правила над другими, в случае одновременной применимости нескольких правил к одному фрагменту входного предложения. Т.е. если к одному фрагменту можно применить правило 1 и правило 2, то будет выбрано, то правило, которое имеет больший приоритет. В случае равенства приоритетов будет порождено два варианта разбора входного предложения, каждый из которых будет содержать уникальный фрагмент, на котором произошло пересечение правил.

В списке блокируемых правил указывается имя того правила, для которого мы хотим, чтобы данное правило блокировало другие правила на том фрагменте, где оно применилось. Если правило заблокировано на определенном фрагменте, то оно больше не может быть к нему применено вне зависимости от его приоритета. Необходимость блокировки была обнаружена в ходе тестирования системы фрагментарного анализа и была вызвана тем, что один лишь приоритет не обеспечивает должного контроля над применением правил коллекции.

Коллекция правил

Для каждого естественного языка, для которого должен выполняться фрагментарный анализ, формируется свой набор правил по выделению фрагментов. Такие правила объединяются в коллекции, по одной на каждый язык. Структура коллекции имеет особый вид, важный с точки зрения работы описываемого анализа.

Структура коллекции правил фрагментарного анализа.

- 1) Наибольший приоритет имеют (а значит, и применяются раньше других) правила снимающие омонимию. Т.е. такие правила, которые, определив фрагмент, удаляют в одном или нескольких его омонимах все слова, кроме одного.

Пример такого правила для английского языка:

Левая часть: that (указательное местоимение) + .(точка).

Правая часть: оставить от 1-ого омонима только указательное местоимение + 2-ой омоним оставить без изменений.

Данное правило снимает со слова that (оно может быть указательным местоимением, союзом, относительным местоимением и наречием) омонимию, если оно употреблено в составе фрагмента, где сразу за ним идет знак точки, т.е. в конце предложения.

Еще один пример подобного правила:

Левая часть: любой артикль + любое слово, которое может быть существительным, глаголом в форме инфинитива, прилагательным.

Правая часть: 1-ый омоним оставить без изменений + из 2-ого омонима оставить существительное или прилагательное.

В английском языке многие существительные могут быть и глаголами, что резко увеличивает количество вариантов при грамматическом анализе, однако, если перед омонимом стоит артикль, то можно однозначно сказать, что данное слово здесь не может быть глаголом.

- 2) Далее расположены правила выделяющие потенциальные границы оборотов, фраз и простых предложений. Границей оборота и простого предложения считаются: знаки препинания (такие как точка, запятая, точка с запятой, тире, двоеточие, знаки восклицания и вопроса), союзы That и If, а также слова WH-группы (Who, Which, What, When). Для всех этих слов написаны правила помечающие их соответствующими метками.

В этом же блоке приоритетов выполняются правила определения сказуемых в предложении. Например, омоним считается сказуемым, если он состоит из одного слова – глагола. Каждое правило, идентифицировавшее фрагмент как сказуемое, выделяет его с помощью метки.

3) Наиболее низкий приоритет имеют правила, маркирующие обороты, фразы и простые предложения, а также разделяющие их по типам. Такая маркировка происходит на основе меток границ полученных в результате применения правил, описанных в пункте 2.

Следует отметить, что определение конца фрагмента представляет собой задачу гораздо более сложную, чем определение начала фрагмента. Так как на синтаксический анализ должен поступать фрагмент, который разбирается далее по заранее определенному правилу, задача определения конца фрагмента часто ограничивает область применения фрагментарного анализа. Чтобы распознать конец фрагмента, не заканчивающегося знаками препинания, во многих случаях требуется сложная система правил коллекции фрагментарного анализа. В таких ситуациях анализ наиболее эффективно осуществлять с помощью коллекции правил блока синтаксического анализа.

Алгоритм работы модуля фрагментарного анализа

После полного описания правил и их коллекций можно описать полный алгоритм работы всего модуля. На вход фрагментарного анализа поступает предложение, состоящее из омонимов для каждого слова, в которых определена часть речи и грамматические параметры. На каждом шаге работы фрагментарного анализа для каждого правила из коллекции для данного входного языка делается попытка найти в предложении фрагменты, к которым оно применимо. После этого происходит анализ потенциально применившихся правил и, в случае пересечения позиций их применений, из них выбираются те, что имеют наивысший приоритет. Далее избранные на предыдущем шаге правила применяются, в результате чего получается новое входное предложение (или предложения) содержащее меньшее количество слов в омониме и/или дополнительные метки, выделяющие те или иные фрагменты. Это новое предложение (или все предложения поочередно) подаются опять на вход фрагментарного анализа, где к ним применяются приведенные выше операции. Процесс повторяется до тех пор, пока ни одно правило из коллекции не сможет быть применено к входному предложению.

Результатом работы фрагментарного анализа является предложение, в котором сокращена омонимия относительно исходного и расставлены метки, маркирующие желаемые составителем коллекции правил части предложения

(такими частями может быть простое предложение, фраза, оборот, сказуемое и т.п.).

Выводы

Представленный в статье модуль фрагментарного анализа выполняет две основные функции: снятие омонимии и маркировка заданных фрагментов входного предложения.

Возможность снятия омонимии позволяет сократить количество вариантов разбора предложения грамматическим анализом, что серьезно ускоряет обработку предложения и работу системы в целом.

Метки, расставленные над фрагментами входного предложения, позволяют идентифицировать его структуру, которая может быть эффективно использована в ходе синтаксического разбора предложения. Это позволяет увеличивать глубину разбора синтаксического анализа для сложных предложений. В некоторых случаях, в особенности для длинных предложений со сложной структурой, корректный разбор предложения одними только правилами синтаксического анализа практически не осуществим из-за слишком большого числа деревьев разбора. Введение модуля фрагментарного анализа, использующего информацию о структуре отдельных типичных фрагментов переводимого текста, позволяет значительно расширить спектр корректно разбираемых предложений, существенно экономит требуемый ресурс оперативной памяти и время машинного перевода, а также в целом ощутимо повышает качество перевода.

Список литературы

1. Collins. Collins Cobuild English Grammar. / Harper Collins. 1992.
2. Vilson J. Leffa. Clause processing in complex sentences. / In Proceedings of the First International Conference on Language Resource&Evaluation, May 1998, volume 1, pages 937 – 943.
3. Radolf Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. A Comprehensive grammar of the English Language. / Longman. 1985.
4. Georgiana Puscasu. A Multilingual Method for Clause Splitting. / University of Wolverhampton. 2003.

