

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ СПОРТИВНОГО КОММЕНТАРИЯ

AUTOMATIC GENERATION OF SPORTS COMMENTARY

М.Ю. Токарева (mashka@ecfor.ru)

Е.И. Большакова (bolsh@cs.msu.su)

Е.А. Бордаченкова (lenabord@mail.ru)

МГУ им. М.В.Ломоносова, факультет ВМиК

Описывается метод автоматического синтеза связного текста для оперативного комментирования динамичных спортивных соревнований. Ключевые моменты – гибкий выбор очередного события для комментирования и построение для него строки комментария на основе подходящего шаблона фразы. Кратко характеризуется разработанная система автоматического комментирования гонок «Формула-1».

Введение

Задача генерации текста на естественном языке является одной из наиболее сложных задач искусственного интеллекта и компьютерной лингвистики. Существует большое количество подходов к синтезу текста на естественном языке [1]. Малоисследованной областью является динамическая генерация текста, или генерация в режиме реального времени [2, 3] – так называется генерация текста, происходящая параллельно с выполнением некоторого процесса. Временные ограничения делают невозможным применение традиционных методов синтеза текста, требующих использования обширных лингвистических знаний и достаточного большого количества времени. В реальном времени требуется гарантия того, что генерация текста не помешает вовремя реагировать на изменение ситуации, и это вынуждает выбирать простые методы, позволяющие достичь компромисса между скоростью генерации и качеством выходного текста [3].

Примерами областей, где необходима динамическая генерация текста, является создание навигационных систем реального времени для пешеходов (главным образом, для слепых людей) и для автомобилей, систем слежения движения транспорта (например, прибытия автобусов или самолетов), а также систем комментирования спортивных соревнований.

Известно несколько экспериментальных систем, разработанных для автоматической генерации устного комментария к спортивным соревнованиям, например, система Mike [4] для комментирования футбольных матчей и система комментирования парусных регат [5]. Большинство систем спортивного комментирования использует для генерации ту или иную разновидность метода шаблонов. Ориентация на генерацию устной речи приводит к необходимости создания механизмов, прерывающих устное изложение текущего комментария в случае, когда необходимо отреагировать на внезапно возникшее чрезвычайно важное событие. Разработке таких механизмов и уделялось наибольшее внимание в вышеупомянутых системах генерации. Вопросы же обеспечения связности и языковой разнообразности комментария были изучены недостаточно.

В настоящей работе описывается разработанный метод автоматического синтеза связного текста для оперативного комментирования спортивных соревнований. В качестве предметной области были выбраны гонки «Формула-1». Такой выбор был мотивирован тем, что гонки – динамичные соревнования с несколькими участниками и разнообразными событиями. Более того, несколько событий могут произойти одновременно. Поэтому временные рамки, за которые то или иное событие может быть прокомментировано, ограничены. Указанные особенности потребовали разработки ряда средств, обеспечивающих как своевременность, так и связность генерируемого комментария.

Ключевые моменты описываемого метода – гибкий выбор очередного события для комментирования и построение для него строки комментария на основе шаблона фразы, подбираемого из нескольких применимых шаблонов, с учетом предыстории событий. Использование шаблонов мотивировано тем, что количество различных типов событий соревнования невелико, каждое событие отражается в комментарии одним или двумя предложениями, а время на синтез предложения ограничено.

На основе разработанного метода была создана система автоматического комментирования гонок «Формула-1». Для моделирования гонки и порождения входных данных, используемых системой комментирования, был разработан и реализован симулятор гонок «Формула-1». Система автоматического комментирования генерирует связный текстовый комментарий в ответ на динамически возникающие события в моделируемой гонке. Испытания, проведенные с системой, подтверждают применимость метода для построения текстов, адекватно описывающих динамичные спортивные соревнования.

1. Основные особенности спортивного комментирования

Анализ типичных комментариев спортивных соревнований показывает, что в них выделяются логически слабосвязанные фрагменты следующих видов:

- событийный комментарий,
- фонový комментарий,
- оценочный комментарий.

Событийный комментарий заключается в описании происходящих событий и является основной частью комментария спортивного соревнования. Быстрое изменение ситуации не позволяет комментатору подробно освещать каждую деталь происходящего, его задача – успеть сообщить основную информацию до наступления новых событий. Если несколько событий происходит одновременно, то комментируется наиболее важное событие.

Фоновый комментарий содержит информацию общего характера об истории данного вида спорта, правилах проведения соревнований и т.п. Когда в соревновании наступает затишье, комментатор заполняет паузу фоновым комментарием, который прерывается при изменении ситуации в соревновании.

Оценочный комментарий состоит из статистических сведений, собранных по ходу соревнования (например, данных об относительной активности участников соревнования, количестве атак или штрафных санкций) и выдается с некоторой периодичностью, опять же при относительно стабильной ситуации в соревновании.

Основную сложность при автоматической генерации представляет событийный комментарий ввиду ограничений на время синтеза фраз: в отличие от фонового комментария, эти фразы не могут быть полностью заготовлены заранее. Кроме того, часто возникает необходимость описывать событие, не дожидаясь его полного завершения, т.е. фактически по очереди описывать его подсобытия.

В целом, чтобы комментарий, генерируемый для динамичного спортивного соревнования, был близок по качеству к комментарию профессионального комментатора, он должен удовлетворять следующим требованиям:

Адекватность (ситуации): текст должен освещать в первую очередь события, наиболее важные из происходящих в текущий момент, а также включать фоновый и оценочный комментарий;

Своевременность: основные события соревнования должны быть прокомментированы вовремя;

Разнообразие: для каждого типа события должно быть несколько возможных фраз-вариантов комментирования;

Связность: комментарий должен быть целостным описанием происходящего соревнования, а не разрозненным набором фраз.

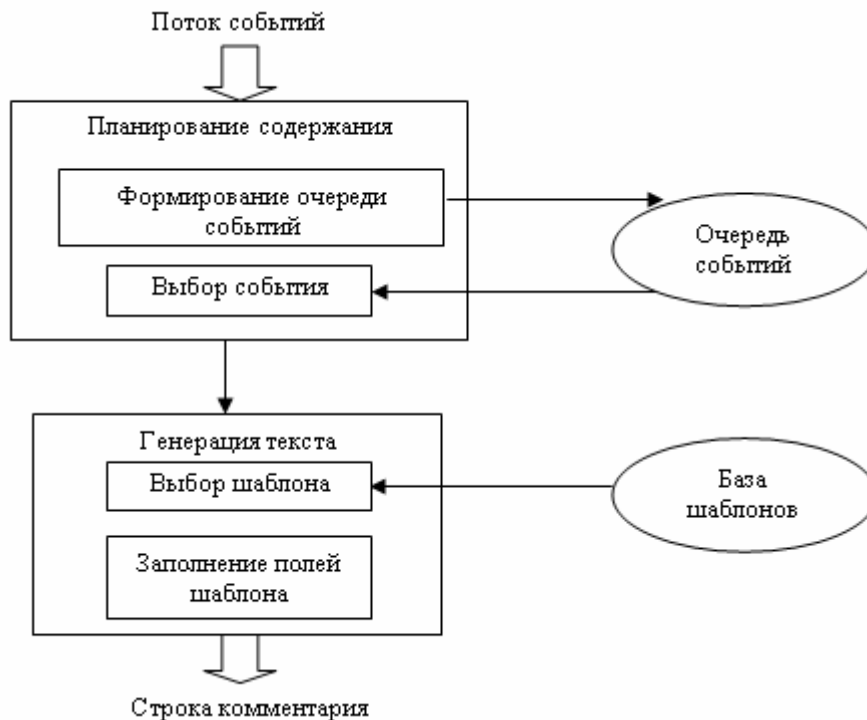


Рис. 1. Схема генерации комментария

2. Схема генерации комментария

Будем рассматривать спортивное соревнование как происходящую во времени последовательность событий, каждое из которых касается определенных объектов, в том числе – участников соревнований. Каждое событие характеризуется типом и атрибутами. Например, в предметной области гонок «Формула-1» одним из

типов событий является *СТОЛКНОВЕНИЕ* с атрибутами: участники столкновения, место столкновения, время начала события. Некоторые типы событий являются взаимосвязанными, например, они представляют собой подсобытия одного протяженного по времени события. Так, в гонках ситуация, когда одна машина обгоняет другую, складывается из подсобытий: *ПОПЫТКА ОБГОНА* и *ЗАВЕРШЕНИЕ ОБГОНА* или *НЕУДАЧНАЯ ПОПЫТКА*.

Своевременность комментирования обеспечивается пошаговым механизмом планирования содержания. По входному потоку событий (упорядоченному по времени возникновения) формируется *очередь событий*. События поступают в очередь гораздо быстрее, чем синтезируется соответствующий комментарий и невозможно успеть прокомментировать абсолютно все события. Поэтому на каждом шаге из очереди выбирается наиболее актуальное событие – см. рис. 1.

Формирование и поддержание очереди событий и выбор из нее актуального события для комментирования составляют этап *планирования содержания*.

На следующем этапе происходит *генерация комментария* для выбранного события. Для каждого типа события заготовлено несколько альтернативных шаблонов фраз. По определенным правилам, с учетом предыстории событий, происходит подбор шаблона и его настройка в соответствии с атрибутами события. Множество шаблонов фраз, необходимых для комментирования соревнования, хранится в *базе шаблонов*. Предыстория событий включает информацию о событиях, которые происходили с каждым участником соревнований.

После того, как событие прокомментировано, оно удаляется из очереди. Если на очередном шаге выбора события для комментирования очередь событий оказывается пуста, выдается фоновый комментарий из заготовленного заранее текста.

Каждое событие в очереди характеризуется *приоритетом* и *временем жизни*, которые устанавливаются при занесении события в очередь.

Приоритет отражает относительную значимость событий. В общем случае приоритет зависит от типа события; например событие *ВЫЛЕТ С ТРАССЫ* имеет более высокий приоритет, чем событие *ПЕРЕТОРМАЖИВАНИЕ*;

значений атрибутов события; например, если события одного типа произойдут с гонщиками, идущими на первом и пятнадцатом месте, то событие, произошедшее с первым, получит более высокий приоритет;

последнего прокомментированного события: если последним было прокомментировано событие, произошедшее с каким-либо участником, то новому событию, имеющему отношение к этому же участнику, присвоится более высокий приоритет.

Время жизни соответствует времени, в течение которого событие остается актуальным. С течением времени актуальность любого события постепенно уменьшается, а некоторые события не имеют смысла комментировать вообще. Например, если гонщик перетормозил в повороте, то комментировать это в следующем повороте уже бессмысленно. Другие события, напротив, не устаревают довольно долго. Как правило, время жизни более значимых событий больше, чем менее значимых. По истечении своего времени жизни событие удаляется из очереди, как ставшее неактуальным.

Приоритеты и время жизни событий подбираются эмпирическим путем для каждой конкретной предметной области. Использование приоритетов и времени жизни событий в очереди позволяет:

- комментировать только актуальные события;
- последовательно комментировать взаимосвязанные события;
- генерировать оценочный и фоновый комментарий.

Если для комментирования было выбрано подсобытие протяженного во времени события, то наивысший приоритет присваивается подсобытию, являющегося продолжением этого события. Таким образом, протяженное во времени событие начинает комментироваться до его полного завершения, и соответствующие предложения генерируются последовательно.

Генерация оценочного комментария организуется следующим образом: с некоторой периодичностью генерируются и помещаются в очередь события специального типа с малым приоритетом и большим временем жизни. Когда в очереди не окажется более важных событий, будет выбрано это событие, и в синтезируемый комментарий будет включена соответствующая информация. Например, для гонок «Формула-1» таким способом сообщается рекорд времени лучшего круга в гонке или количество машин на трассе.

3. Построение строки комментария

Выбор из нескольких шаблонов фраз позволяет разнообразить комментарий однотипных событий. Например, для комментирования события *ОБГОН* из области гонок можно использовать шаблоны

«#PROB:0.5 [Удачно,]#?xI **выходит на ?P место**
«И ?xI **побеждает в борьбе за ?P место**»

Шаблон состоит из фиксированных фрагментов текста и полей, в которые подставляются конкретные значения, когда шаблон применяется к некоторому экземпляру события. В приведенном выше примере в поле ?xI будет подставлено имя обгоняющего гонщика в именительном падеже, в поле ?P – позиция, которую он

стал занимать в гонке, запись #PROB:0.5[**Удачно**,]# означает, что с вероятностью 0.5 в предложение будет включено слово **Удачно**.

Каждому шаблону сопоставлено *условие применимости*. При подборе шаблона из всего набора шаблонов, ассоциированных с событием, отсеиваются те, у которых не выполняются условия применимости, а из оставшихся выбирается один, причем таким образом, чтобы вероятность выбора одного и того же шаблона несколько раз подряд была минимальна.

Условия применимости позволяют при выборе шаблона учитывать влияние события на ход соревнования. Например, в комментировании гонок при наступлении события *ВОЗВРАТ НА ТРАССУ* используется шаблон

«?xI возвращается на трассу, потеряв ?N позиция»

при условии, что позиция гонщика ?xI изменилась не более, чем на 5 мест, и шаблон

«Сейчас ?xI откатывается достаточно далеко назад»,

если позиция гонщика значительно ухудшилась.

Использование условий применимости способствует также улучшению связности текста, позволяя выбирать шаблон в зависимости от предыстории возникновения события. Например, при обработке события *УСПЕШНЫЙ ОБГОН* используется шаблон

«?xI обходит ?yV»,

если количество предпринятых попыток обгона меньше некоторого числа, в противном случае выбирается шаблон

«После упорной борьбы ?xD удается, наконец, опередить ?yV».

В шаблонах имеются поля следующих типов:

- Именное поле – в это поле вставляется имя участника соревнования или название команды; в шаблоне указан нужный падеж имени.
- Числовое поле – оно заполняется информацией о количестве участников соревнования, о позиции участника и т.п.
- Модифицирующее поле – в это поле с заданной вероятностью подставляется фиксированный фрагмент текста, если истинно указанное для поля условие подстановки.

При заполнении полей первых двух типов используются соответствующие атрибуты события. При заполнении числовых полей существительное, идущее за числительным, согласуется с ним в числе и падеже.

Использование модифицирующих полей позволяет по одному шаблону строить предложения разной структуры и тем самым избежать однообразия при генерации текста. Проиллюстрируем это на примерах.

Пример 1. Рассмотрим шаблон

«?xI #LEADER:PROB:0.7# останавливается в боксах».

По условию #LEADER# проверяется, является ли ?xI лидером гонки, и если является, то с вероятностью 0.7 будет сгенерирована фраза:

«Шумахер, лидер гонки, останавливается в боксах»,

в противном случае мы получим строку:

«Шумахер останавливается в боксах».

Следующий пример иллюстрирует использование предыстории событий при настройке шаблона: если некоторое событие повторится вновь с одним и тем же гонщиком, то это будет отображено в комментарии.

Пример 2. Шаблон

«У ?xR #REPEAT# загорелся мотор!»,

дает возможность получить предложение

«У Монтойи проблемы продолжают: загорелся мотор!»

или

«У Монтойи загорелся мотор!»

в зависимости от того, были ли у данного участника неудачи ранее.

Обычно в связном, стилистически грамотном тексте отсутствуют неоправданные повторы, для чего используются различные способы именования субъекта. Специальная процедура (генератор именуемых выражений) в зависимости от контекста генерирует личное местоимение, имя и/или фамилию в нужном падеже. Способ именования участника события зависит от следующих факторов:

- как долго участник не упоминался в тексте комментария соревнования;

- какими языковыми средствами участник именовался в последней сгенерированной фразе;
- был ли он активным или пассивным членом предложения в уже построенном комментарии.

Примером работы генератора именуемых выражений может служить следующий фрагмент комментария:

*«Баррикелло **пытается пройти** Клина **на входе в поворот**.*

*Кристиан Клин **перекрещивает траектории**»*

При генерации второго предложения в именное поле шаблона были подставлены имя и фамилия гонщика, так как в предыдущей фразе он являлся пассивным членом предложения.

4. Система комментирования гонок «Формула-1»

Описанный метод синтеза спортивного комментария был реализован в системе автоматического комментирования гонок «Формула-1». Предварительно был произведен анализ реальных комментариев и специализированных печатных материалов, и на этой основе было выделено более тридцати типов событий, возникающих в гонках (прокол шины, занос, заезд на питлейн, поломка спойлера и т.д.), а также была создана база, содержащая более ста шаблонов типичных фраз.

Также был разработан и реализован модуль (симулятор гонки), который моделирует гонку и генерирует входной поток событий для системы комментирования. Симулятор гонки осуществляет перемещение машин по трассе, моделирует принятие решений пилотами и взаимодействие машин, фиксирует события, произошедшие в гонке. Численные характеристики моделирования (время между наступлением событий и их длительность, количество происшествий с каждым участником) близки к соответствующим показателям реальных соревнований.

Интерфейс системы включает в себя отображение гонки в двумерной графике, список гонщиков и строку текущего комментария, которая отображается на экране достаточное для прочтения время (порядка двух секунд).

В таблице 1 приведен небольшой фрагмент построенного комментария, позволяющий оценить его качество. В левом столбце указано время (минуты, секунды и тысячные доли секунды), прошедшее от начала гонки, в правом столбце – сгенерированный текст. Среднее время генерации строки комментария – 0.01 секунды на процессоре Celeron 2000 MHz.

Заключение

Описан метод автоматической генерации комментария спортивного соревнования. Комментарий состоит из оперативного описания событий соревнования и фонового текста общего характера. При построении строк комментария используются шаблоны фраз, выбираемые из базы шаблонов.

Своевременность, разнообразность и связность синтезируемого комментария обеспечивается взаимосвязанным использованием:

- приоритетов и времени жизни событий;
- предыстории событий;
- условий применимости шаблонов;
- модифицирующих полей шаблонов;
- различных именуемых выражений.

Метод был реализован и опробован в системе генерации комментария для гонок «Формула-1». Испытание программной системы дает возможность заключить, что предложенный метод позволяет строить разнообразные связные тексты, не нарушая временные требования к генерации текста для области, характеризующейся быстрой сменой событий.

Ч:МИН:СЕК:МС	КОММЕНТАРИЙ
0:00:34.740	Физикелла срезает шикану!
0:00:37.140	И пропускает вперед Фернандо Алонсо
0:00:40.740	Баумгартнер срезает шикану!
0:00:42.540	У него проблемы: сломан передний спойлер!
0:00:44.340	Жолт Баумгартнер медленно едет на пит-стоп!
***	***
0:01:44.430	Пиццония вылетает с трассы!
0:01:47.650	Он возвращается на трассу на прежней позиции
0:01:50.070	Баумгартнер останавливается в боксах для ремонта
0:01:51.870	Попытка атаки со стороны Рубенса Баррикелло
0:01:55.350	Но Физикелла не уступает ему
0:01:58.080	И вновь атака на Джанкарло Физикеллу со стороны Баррикелло
0:02:02.040	Удачно, Рубенс выходит на 11 место
0:02:17.250	Шумахер атакует Ряйккёнена
0:02:24.390	И Михаэль побеждает в борьбе за 2 место

Таблица 1. Пример работы системы

Разработанный метод может быть применен при построении систем автоматической генерации текста для предметных областей, в которых динамичное изменение ситуации представимо в виде событий. Для этого надо определить характерные типы событий, подобрать для них приоритеты и времена жизни и разработать соответствующую базу шаблонов фраз. Построенные же по описанному методу процедуры планирования содержания (формирование очереди событий и выбор из нее актуального) и генерации текста (выбор шаблона и его настройка) являются проблемно-независимыми.

Литература

1. *Соколова Е.Г., Болдасов М.В.* Автоматическая генерация текста на ЕЯ (портрет направления) // Труды международной конференции "Диалог 2004". М.: Наука, 2004. С. 565-572.
2. *Cole R., Mariani J., Uszkoreit H., Battista Varile G., Zaenen A., Zampolli A.* Survey of the state of the art in human language technology. Cambridge: Cambridge University Press, 1998. 553 p.
3. *McRoy S., Channarukul S., and Ali S.* Creating natural language output for real-time applications // Intelligence: New Visions of AI in Practice. 2001. P. 21-34.
4. *Tanaka-Ishii K., Hasida K., Noda I.* Reactive content selection in the generation of real-time soccer commentary // Proceedings of COLING-98. San Francisco, USA: Morgan Kaufmann Publishers, 1998. P. 1282-1288.
5. *Esuli A., Cisternino A., Pacini G., Simi M.* Adaptive real time comment generation for sail racing scenarios // Adaptive hypermedia and adaptive web-based systems: Proceedings of second International Conference, AH 2002. Heidelberg, Germany: Springer-Verlag, 2002. P. 392-396.