Пуристическая стратегия и концентрический принцип автоматического моделирования языковых способностей[1]

Крылов С. А.
krylov@rinet.ru

В докладе обосновывается некоторый (автор смеет надеяться, что оригинальный) подход к автоматическому моделированию языковых способностей, который можно охарактеризовать как концентрико-пуристический подход (далее КонПуП). Основная особенность КонПуП – совмещение пуристической стратегии с концентрическим принципом.
Пуристическая стратегия, в отличие от противоположной ей толерантностной стратегии означает ориентацию на создание систем, действующих безошибочно. Пуристическая стратегия предполагает, что те речевые ошибки, которые работающая система допускает на практике, если и допустимы, то лишь в процессе отладки системы, а не в процессе работы тех продуктов, которые предлагаются вниманию пользователя. На практике системы, относящиеся к классу "автоматических словарей" (АСл), обычно следуют пуристической стратегии (в том смысле, что обычно каждая новая версия создаваемого и расширяемого электронного словаря удовлетворяет критерию безошибочности). Между тем реально наличествующие системы автоматического перевода (АП) этим свойством не обладают.
Концентрический принцип (КонП) означает, что в последовательно построенной цепочке "версий" некоторой действующей модели (1-я, 2-я, …, N-1-я, N-я, N+1-я…) на любой стадии её последовательного улучшения то новое, что вносит N-я версия в работу модели по сравнению с N-1-ой, описывает более периферийные слои языковой системы, нежели N-я версия, но более ядерные слои, чем версия N+1-я. КонП предполагает умение создателей модели отличать ядерное в языке от периферийного, а точнее, умение сравнивать два языковых явления по степени их ядерности/периферийности.
На практике степень важности (ядерности-периферийности) языковых знаков оказывается возможным измерить более или менее точным методом. В парадигматическом аспекте самый простой (хотя и несколько грубый) способ такого измерения – это построение частотного инвентаря знаков. В синтагматическом аспекте простейшая (хотя и несколько огрублённая) мера – это длина речевого отрезка, т.е. количество знаков в нём.

I. Предварительные замечания

Проблема принципиальной осуществимости высококачественного автоматического перевода (АП) была поставлена давно, и для её решения предлагались различные пути.

[1] Автор благодарит В. Б. Борщева, М. С. Гельфанда, Г. Е. Крейдлина, А. С. Панину, А. К. Поливанову, В. П. Селегея и Е. Г. Соколову за ценные советы и критические замечания по обсуждаемому вопросу.

Строится правдоподобное рассуждение с целью доказать (по необходимости “нестрого”) оптимистический тезис о достижимости безошибочного АП. Рассуждение носит инженерный характер, т.е. апеллирует не только к логике, но и к здравому смыслу.
Допустим, что уже выбран некоторый исходный (входной) язык А и некоторый переводящий (выходной) язык Б.
Для любого текста (Т1) на языке А:
при условии, что некоторая авторитетная коллегия экспертов-переводчиков (“человеческих”) признаёт данный текст принципиально переводимым (т.е. может усилиями людей-переводчиков построить хотя бы один приемлемый перевод данного текста),
можно за конечное число шагов построить такой переводящий автомат, который способен справиться с этим текстом, т.е. повторить результат “человеческого” перевода, выдав для данного исходного текста в точности тот же набор переводных текстов, которые, по мнению этой коллегии экспертов, являются приемлемыми переводами данного исходного текста.
Уточнение понятий “текст”, “перевод”, “авторитетная коллегия”, “приемлемый” и т.п. не входит в нашу задачу. Попробуем остановиться на выделенных словах (**можно за конечное число шагов построить такой переводящий автомат**) и пояснить, что имеется в виду под этой формулировкой.

II. Обучаемость системы АП как основа осуществимости безошибочного АП (или: о переходе количества в качество)

1. Идея постепенной коррекции строящихся систем АП сама по себе отнюдь не нова. Так, уже на заре машинного перевода (а именно, 21 мая 1958 г. на Первой Всесоюзной конференции по машинному переводу) В. А. Успенский делает вывод о “необходимости сознательного ограничения как грамматического, так и семантического круга текстов, подлежащих переводу” (Успенский 1958/2002: 40/320).
Мысль о заранее запланированном постепенном усовершенствовании системы АП сама по себе безусловно справедлива. Но тезис о сознательном ограничении и о постепенном усложнении может получить и иную конкретизацию.
Установка на простительность стилистических погрешностей на практике привела к тому, что простительными стали считаться любые погрешности вообще – в частности, явные грамматические и смысловые ошибки. При оценке качества АП стало обычным измерять лишь процент ошибок при переводе.
Установка на жанровую ограниченность на практике привела к тому, что задача моделирования естественного языка фактически уступила место задаче моделирования ограниченных (и крайне примитивных) подъязыков отдельных отраслей знания [2].
2. По каким параметрам предлагается ограничить круг переводимых текстов?
АП мыслится как имитация (функциональная аппроксимация) способности человека к “человеческому” переводу. Естественно провести аналогию между онтогенетической последовательностью ступеней обучаемости человека и научно-техническим прогрессом, в ходе которого примитивные версии системы АП постепенно сменяются своими более продвинутыми версиями.
При обучении человека иностранному языку лингводидактика обычно исходит из так называемого концентрического принципа. Грамматика и лексика изучаемого языка вводятся

[2] Сказанное верно не только в отношении научного и научно-технического жанра, но и несколько шире – в отношении большинства деловых жанров вообще. См. подробнее: Шаляпина 1996.

поурочно, причём небольшими порциями. По закону перехода количества в качество постепенное накопление языковых навыков постепенно приводит к овладению языком на высоком уровне. Для разных хронологических этапов обучения человека составляются учебные словари-минимумы, причём словарь-минимум на каждом последующем этапе включает в себя словарь-минимум предыдущего этапа обучения. Множество грамматических конструкций, употребляемых в учебных текстах каждого следующего этапа, также включает в себя множество грамматических конструкций, употребляемых в учебных текстах предыдущего этапа. Вот некоторые статистические закономерности (в учебных текстах для разных хронологических стадий обучения языку):

1. частые слова обычно начинают изучать на более раннем этапе, чем редкие [3].
1. средняя длина предложений в текстах К-го урока чуть меньше, чем в текстах К+1-го урока.
2. чем чаще употребляется в данном языке некоторая граммема, тем раньше её начинают проходить (фактически эту закономерность трудно проверить в тех случаях, когда для данного языка не составлен частотный инвентарь граммем).

3. Бросаются в глаза разительные отличия долговременной памяти человека от долговременной памяти компьютера. Языковые процессоры мгновенно производят морфологический анализ и синтез на основе морфологического словаря в 100 тысяч слов, однако синтаксический анализ и синтез при этом оказывается на порядок хуже, а семантический – почти вовсе отсутствует. Однако трудно представить себе такого иностранца, который бы отлично склонял и спрягал все 100 тысяч слов изучаемого языка, но при этом довольно плохо сочетал бы их в предложения, а лексическими и грамматическими значениями почти совсем не владел бы. Такого информанта мы скорее приняли бы не за иностранца, а в лучшем случае за афатика; но скорее всего, мы усомнимся в его психической нормальности (разумеется, если перед нами не лингвист, изучавший данный язык в сугубо теоретических целях).

При описании многих систем АП в качестве одного из их достоинств упоминается большой объём словаря (без уточнений, за счёт каких пластов лексики достигнут этот объём). Но пополнение словаря малоупотребительной лексикой (например, специальной терминологией) позволяет доводить эту числовую характеристику до любой величины. Между тем хорошо известно, что “покрываемость” текста от такого пополнения будет расти крайне медленно, между тем как включение в словарь самых частых слов языка благотворно сказывается на этой “покрываемости”.

4. Допустим, для исходного языка имеется в наличии грамотно составленный частотный словарь языковых знаков [4] данного языка. В таком словаре каждому знаку приписаны частотные “ранги”: самый частый знак данного языка имеет ранг 1, следующий за ним по употребительности – ранг 2, и т. п.

Назовём инвентарной мощностью системы АП то максимальное число L, которое обладает следующим свойством.

---

[3] О теоретической важности изучения употребительности слов см. Гиндин 1982.

[4] Языковыми знаками корректнее всего считать единицы, сочетание которых подчиняется принципу “композиционности” в смысле Г. Фреге: реально это могут быть морфемы, словоформы, знаки препинания, фразеологизмы. Однако для практических целей (для проведения огрублённого измерения описываемых параметров) проще всего принять условную договорённость считать языковыми знаками (а) сегментные словоформы и (б) знаки препинания, так как их выделение в составе письменного текста наименее затруднительно.

При любом исходном тексте T, знаковый инвентарь (множество знаков) которого целиком вкладывается в подмножество знаков исходного языка, статистический ранг которых не превышает числа L,

тестируемая система переводит этот исходный текст адекватно.

Назовём синтагматической мощностью системы АП то максимальное число S, которое обладает следующим свойством:

При условии, что длина исходного текста не превышает S знаков,

тестируемая система переводит этот исходный текст адекватно.

Назовём "учебным уровнем системы АП по учебнику А" [5] номер U, обладающий следующим свойством:

Для любого исходного текста, который, согласно мнению коллегии из авторитетных экспертов, содержит только знаки (т.е. слова в соответствующих значениях и грамматические единицы в соответствующих значениях), введённые (т.е. пояснённые) в пределах первых U уроков учебника А,

тестируемая система переводит этот исходный текст адекватно.

5. Назовём "достройкой" системы АП такое преобразование её внутренней структуры, при котором ни один из релевантных параметров её мощности (т.е. ни инвентарная, ни синтагматическая, ни учебная мощность) не снижается.

Назовём "элементарным шагом (ЭШ) в увеличении инвентарной мощности" такую достройку системы АП, при которой её инвентарная мощность увеличивается на 1.

Назовём "ЭШ в увеличении синтагматической мощности" такую достройку системы АП, при которой её синтагматическая мощность увеличивается на 1.

Назовём "ЭШ в увеличении учебной мощности" такую достройку системы АП, при которой её учебная мощность увеличивается на 1.

Назовём "общей мощностью" системы АП тройку параметров, включающую "инвентарную мощность" (L), "синтагматическую мощность" (S) и "учебную мощность по учебнику А" (U).

Назовём "ЭШ в увеличении общей мощности" любой из перечисленных трёх разновидностей ЭШ.

6. Правдоподобна гипотеза, согласно которой построение системы АП, обладающей небольшой инвентарной мощностью (например, 100), небольшой синтагматической мощностью (например, 2) и небольшой учебной мощностью (например, 1), является реально осуществимой задачей. И речь идёт не просто о "понятном" переводе, но о переводе, адекватном полностью, т.е. не только не искажающим смысл, но также заведомо соблюдающем нормы выходного (например, русского) языка.

7. Правдоподобна гипотеза, согласно которой любой ЭШ в увеличении общей мощности системы АП является задачей хотя и довольно нелёгкой, но всё-таки реально осуществимой, т.е. что в ходе ЭШ сложность переводческих проблем хотя и будет возрастать, но этот рост будет **контролируемым**.

8. Можно надеяться, что система АП, основанная на стратегии тщательного осуществления ЭШ совершенствования, рано или поздно дойдёт до такого уровня, чтобы, например, правильно переводить примеры из первых нескольких уроков учебника или букваря. Соответственно, система русско-иностранного АП дойдёт до такой степени мощности, чтобы адекватно переводить, например, рассказы из "Азбуки" Л. Н. Толстого.

Это и будет первым ощутимым (неэлементарным) шагом в создании систем безошибочного АП.

---

[5] Напр., для систем англо-русского или русско-английского АП в качестве такого учебника можно взять авторитетный учебник Н. А. Бонк и др.; mutatis mutandis, для любой пары языков существуют более или менее авторитетные практические учебники такого типа.

9. Если каждый ЭШ является реально осуществимой задачей, то по индукции нетрудно увидеть, что рано или поздно (скорее, впрочем, поздно, чем рано) можно построить систему такой общей мощности, что текст любого уровня сложности (при условии, что эксперты по “человеческому” переводу сочтут его переводимым!) окажется “по зубам” такой системе
10. Может показаться странной сама постановка вопроса о том, чтобы через 45 лет после Первой Всесоюзной конференции по машинному переводу (1958) предлагать начинать работу по созданию систем АП “с нуля”, т.е. с систем, обладающих минимальной мощностью.
Однако сама задача построения систем **безошибочного** АП, как мне кажется, ни разу не ставилась вообще. Более того, пока не ставилась задача создания безошибочных систем АП, изложенная программа поэтапного усовершенствования систем вообще не могла осуществляться, так она является осмысленной лишь при условии, если имеется в виду построение **безошибочных** систем АП.
Но с чего-то ведь надо начинать. Лучше поздно, чем никогда!

III. Три пояснительных замечания

1. Пополнение словаря и повышение инвентарной мощности

"Улучшение" процессора означает повышение "общей мощности" процессора. Но далеко НЕ ВСЯКОЕ ПОПОЛНЕНИЕ СЛОВАРЯ влечёт за собой ПОВЫШЕНИЕ ИНВЕНТАРНОЙ МОЩНОСТИ процессора. Ведь если инвентарная мощность процессора равна (сегодня), скажем, W, то пополнение словаря знаком ранга R приведёт к повышению инвентарной мощности исключительно при том условии, что W=R-1. Т.е. при том условии, что наш процессор сегодня уже не делает ни одной ошибки при обработке любого предложения, все знаки которого имеют ранг меньше R.
Так, если тестируемый процессор делает ошибки при переводе предложений, состоящих из весьма употребительных слов (типа "на", "не", "и", "он", "часа", "привет", "до", "дело", "утром", "с", "меня" и т.п.), то пополнение словаря нашего процессора словами типа "проползать", "ревнивый", "теннисистка", "вприкуску", "триллион", "прикасаться", "многоразовый", "антивирусный", "хренотень", "обожествлять", "омоновец" и проч., ДАЖЕ при условии, что они отлично описаны, НЕ ПОВЫСИТ инвентарную мощность процессора. Пока делаются ошибки при переводе относительно УПОТРЕБИТЕЛЬНЫХ слов, инвентарная мощность по определению будет оставаться крайне низкой.

2. О безошибочности

Особенностью предлагаемого подхода к проблеме БЕЗОШИБОЧНОСТИ является то, что безошибочность мыслится как ОБЯЗАТЕЛЬНАЯ ПРЕДПОСЫЛКА создания действующего процессора. Такой процессор должен быть безошибочным не "в идеале", а НА ЛЮБОЙ СТУПЕНИ своего постепенного совершенствования. А значит, СОВЕРШЕНСТВОВАНИЕ (УЛУЧШЕНИЕ) такого процессора состоит не в постепенном "уменьшении процента ошибок", а в постепенном РАСШИРЕНИИ круга текстов, которые он умеет обрабатывать - в частности, переводить (по определению, "умеет" означает "умеет безошибочно").
Утопично ли говорить о безошибочности? Думаю, что нет. И вот почему:

3. Системы типа "автоматический словарь" (АСл) (включая, напр., LINGVO фирмы ABBYY), по определению построены на основе принципа безошибочности и безукоризненно его придерживаются. Т. е. для АСл пуризм является аксиомой.

4. Системы типа "Translation Memory" тоже основаны на принципе безошибочности (по крайней мере, безошибочности, "субъективно оцениваемой" конкретными пользователями).

Цель вышеприведённых рассуждений - ответить на вопрос, какая стратегия работы предпочтительна для того, чтобы "экстраполировать" принцип безошибочности из сферы АСл (где он является аксиомой) в сферу АП (где он пока что не только не имеет статуса аксиомы, но скорее даже напротив, воспринимается как "ересь").

3. Толерантностный и пуристический подход к ошибкам

Разница между традиционным (условно говоря, "толерантностным") подходом к ошибкам процессора и предлагаемым мною подходом (условно говоря, "пуристическим" подходом) состоит в следующем. "Толерантностый" процессор, получив на входе "трудный" текст, ведёт себя точно так же, как если бы он получил на входе "лёгкий" текст: а именно, ПЫТАЕТСЯ обработать его "как-нибудь" и выдать хоть "какой-нибудь" результат (в случае АП - это "какой-нибудь" перевод), причём БЕЗ ГАРАНТИИ КАЧЕСТВА. Между тем "пуристический" процессор, получив на входе "трудный" текст, ведёт себя иначе, чем если бы он получил на входе "лёгкий" текст. А именно, получив "лёгкий" текст, пуристический процессор обрабатывает его и выдаёт результат (в случае АП - это "адекватный" перевод), причём С ГАРАНТИЕЙ КАЧЕСТВА. Однако, встретив "трудный" текст, "пуристический" процессор отвечает пользователю примерно так: "Текст слишком труден, гарантированного результата дать не могу. Советую включить режим гипотетического разбора". А в режиме "гипотетического" разбора, разумеется, никакой гарантии качества не даётся, так как этой разновидностью анализа занимается уже другой - толерантностный - процессор .

Литература

1. Гиндин С. И. Частота слова и его значимость в системе языка // Ученые записки Тартуского ун-та, 1982, вып. 628 (часть серии “Лингвостатистика и вычислительная лингвистика”), с. 22-54.
2. Успенский В. А. Итоги работы секции алгоритмов машинного перевода // Машинный перевод и прикладная лингвистика, № 1(8), с. 31-62.- Перепечатано в кн.: Успенский В. А. Труды по НЕматематике с приложением семиотических посланий А.Н.Колмогорова к автору и его друзьям. Том 1. М.: ОГИ, 2002, с. 314-333.
3. Шаляпина З. М. Автоматический перевод: эволюция и современные тенденции // ВЯ, 1996, № 2, с. 105-117.