ЛОГИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРЕДСТАВЛЕНИЯ ЯЗЫКОВЫХ СТРУКТУР В МАШИННОМ ПЕРЕВОДЕ

LOGICAL AND STATISTICAL METHODS OF LANGUAGE STRUCTURES PRESENTATION IN MACHINE TRANSLATION

Е.Б. Козеренко
Институт проблем информатики РАН
Россия, 117333, Москва, ул. Вавилова, 44 корп. 2
E-mail: kozerenko@mail.ru

Ключевые слова: машинный перевод, трансфер, синтаксис, семантика, функциональное значение, статистические методы, фразовая структура

В работе рассматриваются методы представления языковых структур, относящиеся к различным исследовательским парадигмам. Используемый нами подход базируется на сочетании лингвистических эвристик, логических правил и статистических методов в процессе машинного перевода методом трансфера. Вероятностные расширения вводятся в систему логических правил синтактико-семантического анализа предложений.

1. Введение

Предшествующий период исследований в области машинного перевода характеризовался усиленным вниманием к моделированию внутренних логико-семантических закономерностей языкового строя и функционирования языка, и разработки практических приложений были основаны на эвристических правилах различной степени детализации. При этом методы разрешения неоднозначности языковых структур также были основаны на системах условных правил. В последние несколько лет значительное продвижение в области машинного перевода было достигнуто за счет применения методов, основанных на машинном обучении, вероятностных моделей для анализа языковых структур. Эти подходы, в частности, успешно применялись в речевых системах перевода [1,2].

Еще два важнейших фактора определяют направление развития исследований в области обработки естественного языка на современном этапе: появление больших корпусов параллельных текстов и значительные вычислительные ресурсы современных систем, позволяющие накапливать и использовать ранее переведенные текстовые фрагменты, так называемая «переводческая память», машинный перевод, основанный на прецедентах (Translation Memory, Example-based Machine Translation) [3,4]. Отечественная разработка, использующая принцип переводческой памяти – это система фразеологического перевода [5].

База прецедентных переводов может формироваться вручную – в этом случае перевод будет достоверен, но трудозатраты весьма значительны, а может быть подготовлена на основе больших параллельных корпусов, когда автоматически подбираются шаблоны для перевода [6,7]. В последнем случае необходимы средства редактирования и фильтрации перевода, поскольку неизбежно появляется очень большое число шумов и избыточных недостоверных правил.

Самые последние результаты исследований свидетельствуют о том, что наилучшие показатели точности перевода достигаются расширением традиционных логико-лингвистических подходов методами машинного обучения и использованием корпусной статистики.

2. Методы машинного обучения в обработке естественного языка

Машинное обучение в значительной степени основано на стохастической исследовательской парадигме, уходящей корнями в разработки алгоритмов распознавания речи, символов, исправления орфографии. Базовым методом для решения многих задач, в частности, определения и разметки частей речи, вероятностного грамматического разбора, является правило Байеса. В архитектуре стохастических систем в основном используется алгоритм динамического программирования.

Алгоритмы обучения могут быть двух типов: неуправляемые и управляемые. Неуправляемый алгоритм должен вывести модель, пригодную для обобщения новых данных, которые ему ранее не предъявлялись, и этот вывод должен быть основан только на данных. Управляемый же алгоритм обучается на множестве правильных ответов на

данные из обучающей выборки, таким образом, что выведенная модель даст более точные решения.

Целью машинного обучения является автоматический вывод модели для некоторой области на основе данных из этой области, таким образом, система, обучаемая синтаксическим правилам, должна быть обеспечена базовым набором правил фразовых структур. В последнее время больше внимания исследователей было уделено построению N-граммов, отражающих сложности синтаксических и семантических структур [8,9], применению N-граммов переменной длины [10], включению семантической информации в N-граммы, например, семантических ассоциаций слов, основанных на латентном семантическом индексировании [11].

Статистические методы обработки естественного языка расширяют схему основных существующих подходов к машинному переводу – прямого перевода, переноса (трансфера), и подхода на основе языка-посредника (интерлингвы) [12]. Машинный перевод на основе статистики был впервые предложен в [13,14].

Отправная точка для любой системы обработки естественного языка – проектирование модуля определения и разметки частей речи (тэггера). Различные стохастические тэггеры появились в 1980-е годы. Общая идея всех стохастических тэггеров заключается в выборе наиболее вероятного тэга (т.е. частеречной метки) для данного слова. Чаще всего для вероятностных тэггеров используются Марковские модели, так, например, для некоторого данного предложения или последовательности слов выбирается последовательность тэгов, которая максимизирует следующую формулу:

P(слово| тэг)*P(тэг|предыдущие n тэгов).

Еще один подход к машинному обучению, основанный на правилах и стохастическом тэггировании (разметке частей речи), известен как обучение, основанное на трансформациях (Transformation-Based Learning, TBL). TBL – это метод управляемого обучения с использованием заранее размеченного обучающего корпуса.

Для вероятностного грамматического разбора применяются стохастические грамматики.

- Вероятностная контекстно-свободная грамматика, ее определение - G = (N,T,P,S,D), где N – это множество нетерминальных символов, T – множество терминальных символов, P – множество продукций вида A -> b, где A – это нетерминальный символ, b – это цепочка символов, S – специальный исходный симвло, D – это функция, приписывающая значения вероятности каждому правилу из множества P.
- Вероятностная грамматика замещения деревьев: ее определение то же, что и для вероятностной контекстно-свободной грамматики, но здесь мы имеем дело скорее не с правилами, а фрагментами деревьев произвольной глубины, верхние и внутренние узлы которых – нетерминальные символы, и листья которых являются терминальными и нетерминальными символами, при этом значения вероятности приписываются этим фрагментам. Таким образом, вероятностная грамматика замещения деревьев является обобщением вероятностной контекстно-свободной грамматики, при этом более мощной, поскольку можно приписывать значения вероятности фрагментам или даже целым схемам разбора.

Очень важный стимул развития исследований в области обработки естественного языка – это рынок систем машинного перевода, который достиг зрелости в 2002 – 2004 годах. Все большее число компаний и организаций за рубежом и в нашей стране осознает преимущества реализации и использования систем машинного перевода, настроенных на задачи в своих сферах деятельности для повышения конкурентоспособности в борьбе за клиентов, говорящих на различных языках.

Нами были проанализированы материалы и описания свыше 180 проектов по машинному переводу с точки зрения базовых моделей лингвистических знаний, степени детализации семантических представлений, методов, используемых для анализа и порождения языка, производительности и функциональности систем, а также доли ручного труда в процессе подготовки правил. Для тех систем, которые доступны в Интернете, использовалась серия тестовых примеров для сравнительного анализа перевода наиболее проблемных языковых явлений.

Наш анализ показал, что современные системы машинного перевода могут быть отнесены к трем группам:

- системы, основанные на статистическом подходе, использующие обучающие наборы данных и параллельные корпуса; в этих системах роль человека в проектировании процессов формирования лингвистических знаний и разрешения неоднозначности сведена до минимума; правила автоматически извлекаются из текстов; также автоматически выявляются контекстные зависимости, на основании которых определяется значение неоднозначных слов или словосочетаний; достоинство этого подхода заключается в значительной или даже полной автоматизации процесса построения базы лингвистических знаний, однако, этот подход значительно осложняется тем, что автоматически выявленные правила часто избыточны, повторяют друг друга, поэтому необходима фильтрация и обобщение правил, что также требует участия человека-эксперта;
- системы, в основе которых лежат детально разработанные человеком правила разбора и

генерации естественного языка, использующие когнитивные модели, в том числе с глубинно-семантическими представлениями; основным ограничением этого подхода является невозможность построения эвристик, предусматривающих все возможные языковые конфигурации и правила разрешения неоднозначности слов и синтаксических структур для всех случаев; в настоящий момент в некоторые из этих систем правил также вводятся вероятностные расширения, призванные частично решить некоторые из указанных проблем;

- системы, которые исходно основываются на двух парадигмах исследования и разработки: логико-лингвистических правилах и стохастических моделях.

Последний подход позволяет оптимально использовать преимущества как традиционных систем, основанных на правилах, так и вероятностных методов, которые применяются для тех классов языковых явлений, которые не могут быть достоверно описаны априорно составленными правилами.

Поскольку структуры естественного языка во многих случаях бывают неоднозначными или многозначными, это приводит к множественности возможных переводов с одного языка на другой. Для разрешения неоднозначности используются вероятностные грамматики разбора, которые предлагают следующее решение: выбор наиболее вероятной интерпретации структуры в данном контексте.

3. Вероятностное моделирование синтаксического разбора предложений

Значения вероятностей для вариантов разбора могут быть получены как на основе корпусной информации, так и на основе лингвистических экспертных знаний. В последнем случае мы имеем дело с достоверной информацией, закрепленной в грамматических системах языков на основе многовековой практики. Значения вероятностей для каждого возможного варианта грамматического разбора (т.е. развертывания нетерминального узла) вычисляются на основе частот встречаемости таких вариантов разбора в существующих текстовых корпусах с синтаксической разметкой (treebanks). Производится подсчет числа раз (N), когда используется некоторый вариант развертывания узла ($\alpha \to \beta$) с последующей нормализацией:

$$P(\alpha \to \beta \mid \alpha) = \frac{N(\alpha \to \beta)}{\sum_{\gamma} N(\alpha \to \beta)} = \frac{N(\alpha \to \beta)}{N(\alpha)} \qquad (1.1)$$

Значения вероятности используются в процессе грамматического разбора. Например, вероятностная контекстно-свободная грамматика (PCFG – Probabilistic Context Free Grammar) и вероятностная грамматика подстановки деревьев (PTSG – Probabilistic Tree Substitution Grammar) присваивают вероятность (P) каждому дереву разбора T (т.е. каждому деривату) предложения S. Эта информация является ключевой для разрешения неоднозначности синтаксических структур. Вероятность каждого возможного дерева разбора T определяется как произведение вероятностей всех правил r, используемых для развертывания каждого узла n в дереве разбора:

$$P(T,S) = \prod_{n \in T} p(r(n)) \qquad (1.2)$$

Вероятность однозначного предложения (т.е. предложения, где нам не надо разрешать неоднозначность) равна вероятности единственного дерева разбора для этого предложения, т.е. P(T,S)=P(T). Вероятность же неоднозначного предложения равна сумме вероятностей всех возможных деревьев разбора ($\tau(S)$) данного предложения:

$$P(S) = \sum_{T \in \tau(S)} P(T,S) = \sum_{T \in \tau(S)} P(T) \qquad (1.3)$$

Вероятность полного разбора предложения вычисляется с учетом категориальной информации для каждой головной вершины каждого узла. Пусть n – синтаксическая категория некоторого узла n, a h(n) – головная вершина узла n, m(n) – материнский узел для узла n, таким образом, мы будем вычислять вероятность p(r(n)|n, h(n)), для этого мы преобразовываем выражение (1.2) таким образом, что каждое правило становится обусловленным своей головной вершиной:

$$P(T,S) = \prod_{n \in T} p(r(n) \mid n, h(n)) \times p(h(n) \mid n, h(m(n))) \qquad (1.4)$$

Грамматика, применяемая для нашей системы правил, - это вероятностная грамматика замещения функциональных деревьев, задающая правила многовариантного когнитивного переноса. В нашей системе грамматики функциональные значения языковых структур определяются категориальными значениями головных вершин. Основой системы логических правил являются обобщенные когнитивные структуры, извлекаемые из систем грамматических категорий некоторых европейских языков и функциональных ролей языковых единиц в предложении.

Вероятностные характеристики вводятся в правила унификационной грамматики в виде весов, присваиваемых деревьям разбора. Наша формальная грамматика основана на правилах фразовых структур, при этом отношения доминирования реализуются через головные элементы, фразовые типы реализуются в иерархических структурах с механизмом наследования атрибутов (задаваемых

явно или по умолчанию), которые позволяют делать обобщения разнообразных типов конструкций. Каждому правилу, задающему отношения зависимости, поставлено в соответствие множество потенциальных линейных структур, в которых реализуются эти отношения.

Неоднозначные и многозначные синтаксические структуры учитываются в многовариантной грамматике когнитивного переноса. В настоящее время на основе функционально-семантического подхода в нашем исследовательском проекте разработана Многовариантная Когнитивная Трансферная Грамматика (МКТГ), учитывающая неоднозначность синтаксических структур, и включающая свыше 300 правил для англо-русского перевода.

МКТГ -правило – это контекстно-зависимая продукция, и деривационный процесс может определяться переходами И/ИЛИ, причем эти два механизма вводят лексическую и структурную неоднозначность, что является центральным свойством естественных языков. При нашем подходе прямое кодирование возможных атрибутов глагольных ожиданий тоже, в основном, осуществляется посредством структур когнитивного переноса. Таким образом, МКТГ является функционально-семантическим вариантом вероятностной грамматики замещения деревьев.

4. Заключение

Подход, основанный на сочетании грамматики когнитивного переноса и стохастических методов, обеспечивает надежную и расширяемую платформу для моделирования межъязыкового синтактико-семантического трансфера (переноса) и может быть применен к большему числу языков (особенно имеющих сходные структуры категориальных атрибутов – значений).

Наши дальнейшие исследования связаны с введением специальных расширений в существующую систему атрибутов-значений, уточнением семантики проблемных головных вершин и глагольных фреймов, развитием механизмов разрешения неоднозначности фразовых структур с помощью вероятностных методов.

Литература

[1] Kay, M., Gawron, J., and Norvig, P. Verbmobil: A Translation System for Face-to-Face Dialog // CSLI; 1992.

[2] Frederking, R., Rudnicky, A.I., and Hogan, C. Interactive speech translation in the DIPLOMAT project // Proceedings of the ACL-97 Spoken Language Translation Workshop, Madrid, ACL, 1997. Pp. 61-66.

[3] Sumita, E. and Iida, H. Experiments and prospects of example-based machine translation // ACL-91, Berkeley, CA, ACL, 1991. Pp. 185-192.

[4] Brown, R.D. Example-based machine translation in the Pangloss system // COLING-96, Copenhagen, 1996. Pp. 169-174.

[5] Система машинного перевода Г.Г. Белоногова: http://www.viniti.ru/russian/vintrans.htm

[6] Nagao, M. A framework of a mechanical translation between Japanese and English by analogy principle // Alick Elithorn and Ranan B. Banerji (eds.), Artificial and Human Intelligence, Edinburgh: North-Holland, 1984. Pp. 173-180.

[7] Sato, S. CTM: An example-based translation aid system // COLING 14, 1992. Pp. 1259-1263.

[8] Rosenfeld, R. A maximum entropy approach to adaptive statistical language modeling // Computer Speech and Language, 1996. N 10, pp. 187-228.

[9] Niesler, T.R. and Woodland, P.C. Modelling word-pair relations in a category-based language model // IEEE ICASSP-99, IEEE, 1999. Pp. 795-798.

[10] Ney, H., Essen, U., and Kneser, R. On structuring probabilistic dependencies in stochastic language modeling // Computer Speech and Language, 1994. N 8, pp. 1-38.

[11] Jurafsky, D. and Martin, J.H. Speech and Language Processing // Prentice Hall, 2000.

[12] Dorr, Bonnie and Nizar Habash. Interlingua Approximation: A Generation-Heavy Approach. // AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.

[13] Brown, P.F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer & P.S. Roossin. A statistical approach to machine translation // Computational Linguistics, 1990. N 16, pp. 79-85.

[14] Brown P.F., S.A. Della Pietra, V.J. Della Pietra and R.L Mercer. The mathematics of statistical machine translation: Parameter estimation // Computational Linguistics, 1993. 19(2), pp. 263-311.

[15] Pollard, C. and Sag, I.A. Head-Driven Phrase Structure Grammar // University of Chicago Press, Chicago, 1994.

[16] Kozerenko, E.B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA. CSREA Press, 2003. Pp. 49-55.