Интерактивное разрешение неоднозначности различных типов в машинном переводе[1]

И.М. Богуславский[1,2], Л.Л. Иомдин[1], А.В. Лазурский[1], Л.Г. Митюшин[1], А.С. Бердичевский[3]

[1] Институт проблем передачи информации РАН
{bogus,iomdin, lazur, mit}@iitp.ru
http://proling.iitp.ru
[2] Мадридский политехнический университет, факультет информатики
igor@opera.dia.fi.upm.es
[3] Московский государственный университет им. М. В. Ломоносова
alexberd@yandex.ru

Аннотация. Описывается модуль интерактивного разрешения лексической и синтаксической неоднозначности, используемый в системе машинного перевода ЭТАП-3. В случаях, когда система не может автоматически выбрать нужное лексическое значение или синтаксическую конструкцию, она просит сделать это пользователя. При разрешении лексической неоднозначности диалог между системой и человеком частично осуществляется на этапе анализа, а частично – на этапе перевода. Для организации диалога статьи рабочих словарей системы снабжаются диагностическими комментариями и примерами, позволяющими выбрать оптимальный вариант. В настоящее время проектируется модуль разрешения синтаксической неоднозначности, доступный внешнему пользователю. Принципиально различаются ситуации внутриязыковой и переводной лексической неоднозначности, требующие человеческого вмешательства на разных этапах.

I. M. Boguslavsky, L.L. Iomdin, A.V. Lazursky, L.G. Mityushin, A.S. Berdichevsky,

Interactive Resolution of Various Ambiguity Types in Machine Translation

Summary. The paper presents the module of interactive word sense disambiguation and syntactic ambiguity resolution used in the ETAP-3. machine translation system. When the system cannot choose the appropriate word sense or syntactic construction automatically, it asks the user to do so. In word sense disambiguation, the man-machine dialogue is partly performed in the analysis phase and partly during the transfer. In order to secure the dialogue, entries of the working dictionaries of the system are supplemented with diagnostic comments and illustrations that enable the user to choose the most appropriate option. At present, a module of syntactic disambiguation understandable by an external user is developed. A clear distinction is drawn between intrinsic and translational ambiguities, which require human intervention at different stages of processing.

Вводные замечания

Предлагаемый модуль интерактивного разрешения неоднозначности предназначен в первую очередь для системы машинного перевода ЭТАП-3 [1,2,5,6,8], разрабатываемой авторами и их коллегами в ИППИ РАН. Мы исходим из того, что основные свойства системы читателю известны, и поэтому ограничимся здесь кратким описанием синтаксического анализатора – центрального компонента системы, устройство которого важно для понимания дальнейшего изложения.

Синтаксический анализатор получает на вход морфологическую структуру (МорфС) обрабатываемого предложения и преобразует ее в дерево зависимостей, узлы которого соответствуют словам предложения, а направленные дуги помечены именами синтаксических отношений (СинтО). Для построения дерева зависимостей из линейной МорфС используются синтагмы – правила, которые описывают минимальные поддеревья, состоящие из двух узлов, связанных некоторым СинтО. Работа анализатора состоит из нескольких этапов. Вначале синтагмы строят все возможные синтаксические связи, используя все виды имеющейся лингвистической информации и материал обрабатываемого предложения. На последующих

[1] Эта работа была поддержана грантом № 05-06-80256 РФФИ, которому авторы выражают искреннюю признательность.

стадиях анализа лишние связи исключаются с помощью ряда фильтров. Если предложение является лексически и/или синтаксически неоднозначным, синтаксический анализатор способен построить несколько синтаксических структур (СинтС), соответствующих различным интерпретациям.

1. Вечная проблема неоднозначности

Несмотря на быстрый прогресс в области систем обработки ЕЯ, проблема разрешения неоднозначности остается камнем преткновения, особенно в тех случаях, когда в задачу системы входит извлечение смысла.. В последнее время много усилий было потрачено на то, чтобы решить эту задачу чисто автоматическими средствами.

С одной стороны, алгоритмы разрешения неоднозначности учитывают все более детальную информацию, обеспечиваемую лексическими и грамматическими ресурсами систем обработки ЕЯ. Именно так работает большинство систем машинного перевода, и ЭТАП-3 не является исключением[2]. Очевидно, однако, что движение в этом направлении имеет естественные пределы, поскольку оно связано с огромными затратами времени и труда. Кроме того, многие случаи неоднозначности нельзя разрешить автоматически в принципе, так как для них существенны экстралингвистические знания, не извлекаемые непосредственно из текста.

С другой стороны, значительный прогресс был достигнут в развитии статистических методов[3]. Этот подход представляется более перспективным; однако показательно, что даже самые мощные статистические системы (см., напр., [14]) имеют уровень эффективности в задаче различения значений для параллельных корпусов не более 75% – цифра впечатляющая, но не достаточная для многих практических приложений. Нам представляется, что полностью автоматизированные процедуры, даже самые эффективные, не могут обеспечить надежного разрешения лингвистической неоднозначности.

2. Интерактивное разрешение неоднозначности: перспективное решение?

В обоих описанных выше методах участие человека в обработке текста ограничивается его предредактированием и постредактированием. Предлагаемый здесь подход предусматривает вовлечение человека-эксперта в обработку текста в самые ключевые моменты процесса анализа или интерпретации. Мы исходим из того, что эксперт должен владеть входным языком, но не обязан знать выходной язык (хотя, конечно, такое знание повредить никак не может). К примеру, таким экспертом может быть автор статьи, желающий перевести ее на язык, которым он не владеет.

Идея интерактивного разрешения неоднозначности (ИРН) была выдвинута четверть века назад. По данным В. Хатчинза [11], системы МП ALPS и Weidner (США) использовали ИРН для английского языка в начале 1980-х гг. В Maruyama *et al*. [13] излагается метод ИРН применительно к японскому языку, К. Буате и Э. Бланшон в Гренобле активно развивают ИРН на материале французского, английского и других языков (МП LIDIA) [7,9,11]. Среди других систем обработки ЕЯ, использующих ИРН, стоит упомянуть также 1) многоязычную систему МП SYSTRAN, 2) систему ALT-J/E (Япония), 3) систему МП UMIST (Манчестер), 4) систему устного и письменного МП группы Spoken Translation (США), 5) систему многоязычного поиска и навигации в Интернете, разработанную DFKI и Университетом земли Саар (Германия).

Работа над первой системой полномасштабного ИРН для русского языка была начата группой ЭТАП в 2002 г., а год спустя началась работа над аналогичной системой для английского языка.

---

[2] Среди существующих в системе дизамбигуаторов - разрешение неоднозначности по ближайшему линейному контексту (предсинтаксический модуль), фильтрация лишних связей, определение наиболее вероятных кандидатов в вершины дерева с помощью эмпирических весов, динамическое присвоение эмпирических весов элементам дерева зависимостей на ранних этапах синтаксического анализа [12]. Разрешению неоднозначности служит также словарная информация о входном и выходном языке: синтаксические и семантические признаки слова, лексические функции и др.

[3] В системе ЭТАП-3 также реализован соответствующий модуль [3].

2.1 Разрешение лексической неоднозначности

Замысел проекта состоит в том, чтобы обеспечить человека, взаимодействующего с системой МП, простыми и ясными диагностическими описаниями неоднозначных лексических единиц, которые могли бы быть ему предъявлены на определенных стадиях обработки текста. Алгоритм анализа был модифицирован так, чтобы любой сделанный человеком выбор отсекал варианты анализа, несовместимые с ним (с возможностью возвращения к исходной ситации, если выбор окажется тупиковым).

Было определено несколько точек процесса обработки текста, когда должно запрашиваться мнение эксперта, а именно: 1) непосредственно перед тем, как синтаксический анализатор приступает к выбору вершины дерева, 2) сразу после проверки всех гипотетических синтаксических связей, построенных синтагмами, 3) непосредственно перед тем, как делается выбор вариантов перевода.

Было отобрано около 20000 русских слов, у которых леммы (или некоторые словоформы) совпадали с леммами (словоформами) других слов, и для них были написаны диагностические комментарии и примеры. Комментарии могут включать: 1) аналитическое толкование значения слова или его существенный фрагмент; 2) маркер части речи, 3) простые синтаксические признаки, 4) синонимы и/или антонимы слова. В расчете на более продвинутых пользователей могут приводиться английские переводные эквиваленты.

Примеры подбираются так, чтобы максимально облегчить идентификацию значения слова. Отметим, при этом, что подобрать для слова контексты, полностью исключающие возможность употребления его омонима/полисеманта, удается не всегда – «контекст определяет лексическую единицу вероятностно, а не абсолютно» (Дж. Лайонз). В таких случаях качественные комментарии приобретают особую важность.

Вся информация записывается в соответствующих статьях русского комбинаторного словаря.

В настоящее время подобная работа осуществляется и на пространстве английского словаря: сформирован список из 20000 неоднозначных английских слов, для которых пишутся различительные комментарии и примеры.

Подчеркнем особо, что метод ИРН реализуется в системе, ориентированной на получение всех возможных вариантов анализа предложения: мы не ограничиваемся одним вариантом, пусть даже наиболее вероятным. Основанием для такого подхода является, в частности, тот факт, что авторы ЭТАПа-3 видят в системе своего рода испытательный стенд для конкретной лингвистической теории Смысл ⇔ Текст [4]: такая система должна учитывать все допускаемые языком интерпретации в максимально возможной мере.

При подобном подходе статистические методы разрешения неоднозначности отступают на второй план. Хотя система располагает целым рядом механизмов, способных подавлять маловероятные интерпретации, мы прибегаем к ним с осторожностью. Точнее говоря, система допускает два режима работы: (а) автоматический режим, при котором вероятностные соображения максимально используются для отсечения менее вероятных интерпретаций на ранних стадиях, и (б) интерактивный режим, позволяющий получить любую адекватную интерпретацию. В этом режиме роль статистических соображений не сводится к нулю, но становится менее приоритетной.

Приведем два примера, иллюстрирующих работу системы МП в интерактивном режиме для обоих направлений перевода.

Начнем с английского предложения

(1) *You can choose either road or this picturesque footpath.*

В автоматическом режиме ЭТАП предлагает перевод *Вы можете выбрать либо дорогу, либо эту живописную тропу.* Легко убедиться в том, что этот перевод неверен: система восприняла слово *either* как часть составного союза *either...or* ‘либо… либо’. Между тем в таком случае исчисляемое существительное *road* ‘дорога’ должно было бы сопровождаться артиклем. В соответствующем правиле анализа, однако, данного требования не оказалось, что и привело к ошибке.

При включенном интерактивном модуле система предложит пользователю выбрать значение *either*, пользуясь диалоговым окном (рис. 1). Разумеется, знакомый с английским языком человек выберет вариант 1, что даст перевод *Вы можете выбрать любую дорогу или эту живописную тропу*. Добавим, что при выборе экспертом варианта 2 структура не будет построена и система перейдет в автоматический режим; наконец, выбор варианта 3 приведет к уже знакомому нам переводу. Таким образом, усилия, затраченные пользователем, вознаграждаются улучшением

качества перевода. Характерно, что интерактивный режим здесь попутно решает и важную задачу компенсации ошибок или недочетов синтаксических правил.

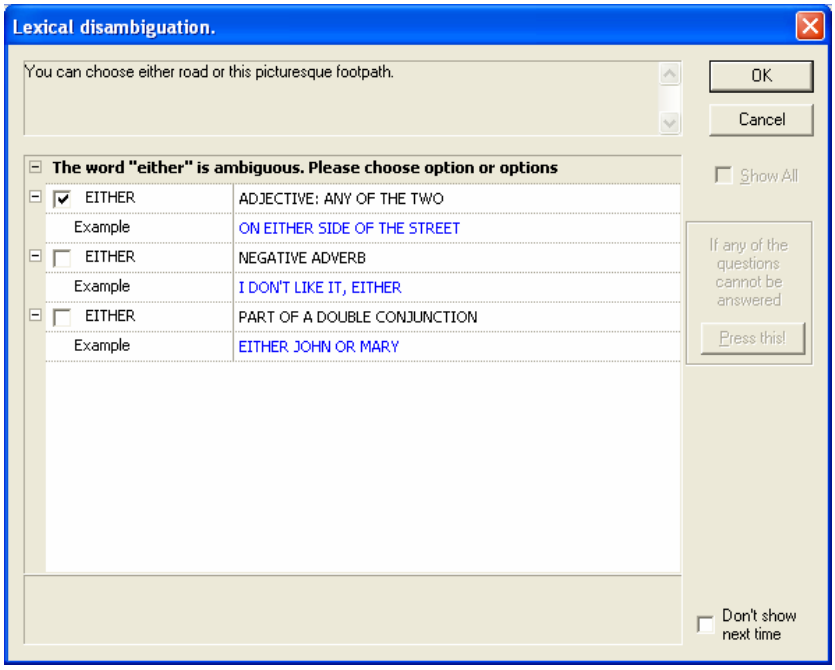

Рис 1. Диалоговое окно системы ЭТАП-3 для разрешения лексической неоднозначности.

Рассмотрим теперь пример разрешения лексической неоднозначности в русском тексте. Хрестоматийное предложение

(2) *Простой солдат вызвал суматоху.*

допускает по меньшей мере две интерпретации: ‘Суматоха была вызвана простым солдатом’ и ‘Суматоха была вызвана простоем солдат’. При работе в автоматическом режиме система ЭТАП-3 поочередно построит для (2) две различные СинтС (рис. 2 и 3), соответствующие этим интерпретациям, и соответственно, породит переводы *The simple soldier has caused a fuss* и *The idle time of the soldiers has caused a fuss*.

Рис. 2. СинтС для первой интерпретации предложения (2)

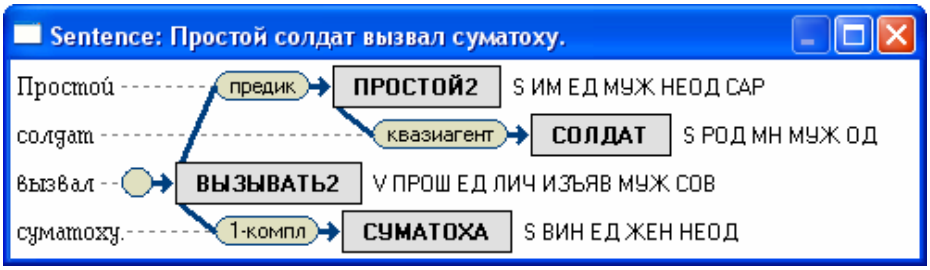

Рис. 3. СинтС для второой интерпретации предложения (2)

В режиме ИРН пользователю будет задан вопрос о значении слова *простой*, и в зависимости от выбранного им варианта: (ПРИЛАГАТЕЛЬНОЕ: НЕСЛОЖНЫЙ) или (СУЩЕСТВИТЕЛЬНОЕ: ВРЕМЯ БЕЗ РАБОТЫ) будет порождена одна из этих структур и соответствующий ей перевод.

Существенно, что модуль разрешения **лексической** неоднозначности здесь помогает справиться с **синтаксической** и **морфологической** неоднозначностью, не задавая пользователю никаких вопросов о синтаксисе или морфологии. Эти «побочные» эффекты возникают регулярно и расширяют возможности лексического модуля ИРН.

Заметим, наконец, что система обработки текстов, вооруженная модулем ИРН, может позволить себе роскошь учитывать редкие омонимы лексических единиц (скажем, английское *see* ‘епархия’) и углубляться в такие тонкости различения значений, которые чисто автоматическая система вынуждена игнорировать из-за угрозы информационного взрыва.

2.2 Разрешение синтаксической неоднозначности

Чтобы система ИРН стала действительно мощным инструментом, она должна иметь средства, позволяющие человеку работать с синтаксисом. Это непростая задача, поскольку рядовой пользователь легко различает лексические значения, но обычно не готов к ответу на синтаксические вопросы.

Уже сейчас ЭТАП-3 предоставляет возможность синтаксического ИРН специалистам, хорошо знакомым с системой. Первоначально модуль разрешения синтаксической неоднозначности предлагал пользователю лишь возможность указать, является ли данное предложение двусоставным или представляет собой именную группу. Это весьма актуально для некоторых типов английских предложений, таких как *Plan changes* ‘изменения плана’ vs. ‘план меняется’ vs. ‘планируйте изменения’ – такие предложения имеют сравнимые вероятности оказаться глагольными и именными. Задавая тип предложения, пользователь влияет на работу правил выбора вершинного узла. В настоящее время этот модуль предлагает диалог, позволяющий пользователю выбирать между синтаксическими гипотезами, имеющими вид бинарных поддеревьев, Такой метод особенно эффективен, если разрешение лексической и синтаксической неоднозначности производится для предложения одновременно.

Результаты, полученные к настоящему времени, являются весьма обнадеживающими. Дальнейшие исследования будут направлены на усовершенствование средств, дающих возможность пользоваться модулем разрешения синтаксической неоднозначности рядовому пользователю, т.е. на разработку наглядного эквивалента использующихся лингвистических формализмов.

В некоторых случаях это достаточно просто. Например, несложным правилом можно эффективно разрешать неоднозначность составляющих, преобразуя неоднозначное предложение *He studies buzzes and whistles* в два более однозначных: *He studies buzzes and he whistles* (≈ он изучает жужжание и свистит) и *he studies buzzes and he researches whistles (*≈ он изучает жужжание и свист), которые и предлагаются на выбор пользователю.

По всей вероятности, разработать универсальные правила наглядного представления произвольных типов синтаксической неоднозначности невозможно. В наши ближайшие планы входит поэтому изучить, какого рода запросы наиболее понятны пользователю, и определить наиболее частотные типы синтаксической неоднозначности, для которых стоит составлять специальные правила наглядного представления. Большим подспорьем может оказаться реализованная в системе ЭТАП-3 система перифразирования.

3. Внутриязыковая и переводная неоднозначность

Важным аспектом нашего подхода является то, что мы отдельно рассматриваем внутреннюю неоднозначность входного языка и неоднозначность, возникающую при переводе. Это различие особенно важно проводить в случае многоязычной системы.

Действительно, если одни случаи неоднозначности не зависят от того, на какой язык переводится текст (и, шире, не зависят от конкретной задачи обработки ЕЯ – примером может служить уже упоминавшийся *простой солдат*), то неоднозначности другого типа возникают только при переводе на определенный язык. Например, при переводе с русского языка на английский слову *свеча* (в прямом значении) соответствует слово *candle*, в то время как при переводе на французский следует учесть материал, из которого сделана свеча, и соответственно

перевести это слово как *la chandelle* (сальная свеча), *la bougie* (стеариновая) или *le cierge* (восковая). Аналогично, не нужно включать ИРН при переводе слова *песня* на английский (*song*), но желательно это сделать при переводе на шведский: *visa* (длинная сюжетная песня типа баллады) vs. *sång* (песня вообще).

Поскольку эти типы неоднозначности имеют разный характер, они разрешаются на разных стадиях обработки предложения: внутриязыковая неоднозначность - во время анализа предложения, а переводная – на этапе собственно перевода[4]. Если бы это различие не учитывалось и оба типа рассматривались одновременно на стадии анализа, то пришлось бы нагрузить описание входного языка информацией обо всех неоднозначностях всех выходных языков, что трудоемко и крайне неестественно. С другой стороны, если бы разрешение внутриязыковой неоднозначности было отложено до стадии перевода, мы лишились бы возможности раннего отсечения неправильных интерпретаций. Насколько нам известно, ЭТАП-3 – единственная система, проводящая четкое различие между указанными типами неоднозначности.

Литература

1. Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Н. В. Перцов, В. З. Санников, Л. Л. Цинман. Лингвистическое обеспечение системы ЭТАП-2. Москва, Наука (1989). 295 стр.
2. Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, В. З. Санников, Л. Л. Цинман. Лингвистический процессор для сложных информационных систем. Москва, Наука (1992), 256 стр.
3. И. М. Богуславский, Л. Л. Иомдин, В. Г. Сизов, И. С. Чардин. Использование размеченного корпуса текстов при автоматическом синтаксическом анализе. // Труды международной конференции «Когнитивное моделирование в лингвистике-2003». Варна (2003), стр. 39-48.
4. И. А. Мельчук. Опыт теории лингвистических моделей класса «Смысл ⇔ Текст» Москва, Наука (1974).
5. Apresian, Ju. D., Boguslavsky, I. M., Iomdin, L. L., Lazursky, A. V., Sannikov, V. Z, Sizov, V. G., Tsinman, L. L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. // MTT 2003, First International Conference on Meaning – Text Theory. Paris, École Normale Supérieure, Paris, June 16–18, 2003, pp. 279-288.
6. Apresjan, Ju. D., Boguslavskij, I. M., Iomdin, L. L., Lazurskij, A. V., Sannikov, V. Z. and Tsinman, L.L. Système de traduction automatique {ETAP}. La Traductique. P.Bouillon and A. Clas (eds). Montréal, Les Presses de l'Université de Montréal. (1993).
7. Blanchon, H. An Interactive Disambiguation Module for English Natural Language Utterances. // Proceedings of NLPRS'95. (Seoul, Dec 4-7, 1995), vol. 2/2: 550-555.
8. Boguslavsky, Igor M., Iomdin, Leonid L., Lazursky, Alexander V., Mityushin, Leonid G., Sizov, Victor G., Kreydlin, Leonid G., Berdichevsky, Alexander S.. Interactive Resolution of Intrinsic and Translational Ambiguity in a Machine Translation System. // CICLing 2005. Lecture notes in computer science. A.Gelbukh (ed.), Springer-Verlag Berlin Heidelberg 2005, pp. 383 - 394.
9. Boitet, C., Blanchon, H. Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup. // Machine Translation, 9/2 (1994), 99-132.
10. Goodman, K. and Nirenburg, S. (ed.). The KBMT Project: A case study in knowledge-based machine translation. Morgan Kaufmann Publishers. San Mateo, California. (1991). 330 p.
11. Hutchins W. Machine translation: past, present, future. Ellis Horwood, Chichester (1986).
12. Iomdin, L. L, Sizov, V. G, Tsinman, L.L. Utilisation des poids empiriques dans l'analyse syntaxique: une application en Traduction Automatique. META, 47. (3). (2002) 351-358
13. Maruyama H., Watanabe, H., and Ogino, S. An interactive Japanese parser for machine translation. // Karlgren, H., ed. Proceedings of the 13th International Conference on Computational Linguistics, v. 2. Helsinki. (1990). 257-62
14. Tufis D., Ion, R., Ide, N. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. // Proceedings of the 20th International Conference on Computational Linguistics, Geneva, August 23–27, 2004, pp. 1312-1318.

[4] В системе ЭТАП-3 результат работы блока анализа – нормализованная СинтС – поступает на вход блока перевода, а затем итог работы данного блока передается блоку синтеза выходного языка. Разграничение между этапами весьма жесткое.