NooJ – НОВАЯ ПРОГРАММА ПО РАЗРАБОТКЕ И ПРИМЕНЕНИЮ ЭЛЕКТРОННЫХ СЛОВАРЕЙ И ГРАММАТИК

Ч. Бланко, Автономный университет Барселоны

Перевод с испанского Евгении Якубович

NooJ (© 2005 Макс Сильберштейн, Университет Франш-Конте) – это, с одной стороны программа по разработке и использованию электронных словарей и грамматик, располагающая, кроме того, мощными средствами маркировки текстовых корпусов и многочисленными функциональными возможностями по вводу и выводу текстов в более сотни форматах (XML включительно). С другой стороны, поскольку программа содержит

встроенные в неё объёмные словари и грамматики более десятка языков (английского, арабского, армянского, болгарского, венгерского, иврита, испанского, итальянского, каталанского, китайского, корейского, латыни, румынского, португальского, французского и т. д.), она является также полной программой по анализу текстов. Оба эти аспекта с пользой могут быть изучены в рамках университетского курса компьютерной лингвистики.

В целом, мы имеем дело с сервисной программой, применимой в тройной перспективе: в разработке и опытном освоении лингвистических ресурсов (перспектива компьютерного лингвиста), в использовании текстовых корпусов (перспектива пользователя, заинтересованного в тщательном изучении текстов, лексическом и стилистическом их анализе) и в преподавании лингвистической инженерии. Следует добавить сюда и четвёртую перспективу: перспективу программиста, заинтересованного в переиспользовании функциональных возможностей NooJ в собственных программах по обработке естественного языка, так как функции NooJ, в его профессиональной версии, доступны в наборе автономных программ Windows (которые могут запускаться через Shell scripts или входить в состав других программ Perl, C++ и т. д.) и в динамической библиотеке .NET, состоящей из классов публичных объектов и методов, которые могут быть переиспользованы в других программах.

NooJ, тем не менее, имеет свою команду пользователей и разработчиков. Крупное международное сообщество объединяет исследовательские группы из 30 с лишним стран, состоящие из лингвистов и программистов, которые занимаются формализацией своих рабочих языков, исходя из очень сходных теоретических оснований, руководствуясь теорией лексики – грамматики, разработанной в своё время в престижной лаборатории Мориса Гросса (Лаборатория документальной автоматики и лингвистики, Paris 7, CNRS). Эта команда выработала огромное количество лингвистических ресурсов, встроенных в NooJ в форме словарей и электронных грамматик четырёх типов: морфологических грамматик (флективных и деривативных), синтактико-семантических грамматик, грамматик перегруппировки вариантов (важнейших при применении в терминографии) и правил устранения двусмысленности.

Таким образом, речь идёт не только о программе по совершентсвованию лингвистической инженерии, но также о точке пересечения, в которой сходятся опыт и знания в области лингвистики, накопленные в течение 15 с лишним лет совместной работы, со времени появления первой версии INTEX в 1992 (основанной на докторской диссертации Макса Сильберштейна, 1989) до полностью разработанной

программы NooJ, способной, например, пошагово осуществить трансформации, описанные З. С. Харрисом, со всей описательной способностью машины Тьюринга.

Несмотря на то, что перед нами недавняя разработка с многочисленными новшествами (что касается последней версии), по ней уже есть специальная литература, которую составляют учебник пользователя, изложенный по-английски самим автором программы (Max Silberztein, NooJ Manual, Université de Franche-Comté 2005), и учебник по введению, написанный на французском языке Мишель Ру (Лаборатория фундаментальной информатики, Марсель). Предусмотрены также упражнения, предназначенные для преподавания в университете, активный список рассылки и группа новостей [http://groups.yahoo.com/group/nooj-info].

Ввиду того что электронные словари под названием DELAS и DELAC, так же как и таблицы лексики - грамматики, были весьма популярны в 90-ые годы в сфере автоматической обработки естественного языка, нам кажется необходимым провести краткое сравнение между вышеупомянутыми словарями и современными словарями NooJ. Словари DELAS описывали простые лексические единицы, то есть единицы, не имеющие отделяющего символа (такого, как пробел или знак препинания). Словари DELAC описывали составные единицы, содержащие этот отделяющий символ (например, *точка зрения, Средние века*). Для обоих словарей нужно было выполнить команду о построении флективных форм перед применением к текстам, в результате чего появлялись флектированные соответствия DELASf и DELACf. И наконец, лексико-грамматические таблицы, несмотря на то, что описывали общие синтаксические характеристики предикатов, использовались в INTEX прежде всего для распознавания идиоматических сочетаний, в которых могли присутствовать вставки (к примеру, *ты витаешь всё время в облаках)*.

Тем не менее, хотя мы ведём речь о весьма приемлемой разработке на базе основательной лингвистической модели, этот тип внутренней организации всё же представлял ряд неудобств. Во-первых, простые и составные единицы искусственно обособлялись в двух различных по своим характеристикам словарях. Эта искусственность усиливалась и тем, что именно теория лексики - грамматики постулировала сходную природу элементов макроструктуры для обоих типов единиц. Во-вторых, сложным оказалось связать отсылками простые и сложные формальные варианты одной и той же лексической единицы (например, программа максимум vs программа-максимум и программа максимум), – проблема, которая обострялась в случае двуязычных словарей, где зачастую выходит так, что эквивалент перевода простой единицы является

составным, и наоборот. Кроме того, автоматическое построение флективных форм составных единиц словаря должно было осуществляться вне программы INTEX. Что же касается таблиц лексики-грамматики, для их реализации необходимо было прибегнуть к метаграфам (внеконтекстуальным грамматикам), с чьей помощью прочитывалось содержание таблиц и двоичные числа, соотносящиеся со свойствами, предложенными для каждого идиоматического сочетания. Как видим, речь идёт о сложном и малодружественном для лексикографа методе.

NooJ способен решить все эти проблемы, используя единственный формат словаря. Словарь NooJ группирует простые и составные единицы, а также идиоматические сочетания в одном и том же типе файла, что не только представляется более удобным, но к тому же позволяет естественным способом устанавливать синонимические отношения или отношения эквивалентности перевода между лексическими единицами, независимо от их принадлежности к определённым частям речи. Построение флективных форм, которое относится и к составным единицам, осуществляется в реальном времени в момент применения словаря к тексту, что исключает необходимость промежуточных файлов DELAF. Нужно также отметить, что эта особенность гораздо больше, чем просто усовершенствование эргономики пользования, так как, в отличие от программы INTEX, которая связывала флективные формы только с соответствующими словарными формами слов – леммами, словарь NooJ позволяет напрямую связывать любую форму флективной или деривативной парадигмы с любой другой формой этой же парадигмы. Это даёт возможность рассмотреть полный спектр вариантов изменения данной фразы. Добавим к сказанному и тот факт, что словарные формы NooJ могут соотносится с "супер-леммой", а именно, единицей, которая не обязательно принадлежит парадигме формы, обнаруженной в тексте, и которая, следовательно, может быть эквивалентом перевода. Таким образом, у нас есть программа, являющая собой подходящую лабораторию для изучения машинного перевода.

Что до маркировки текста, механизм NooJ основывается на системе аннотаций. Аннотация – пара, состоящая из положения и информации, которая соединяет определенное положение в тексте с определёнными величинами, соответствующими свойствам, которые нужно маркировать для каждого положения. Обработав текст, NooJ создаёт набор аннотаций, синхронизированных с текстом, который он сохраняет в TAS (Text Annotation Structure). Сам текст не перезаписывается, а поэтому остаётся доступным в его первоначальной форме для любой последующей обработки.

В том, что касается дружественности, NooJ позволяет в достаточной мере интуитивно разрабатывать сложные электронные грамматики

посредством своей системы графов. Эти графы представляют собой различные логические механизмы, такие как конечные автоматы (AEF) с функцией распознавания цепей; конечные преобразователи (TEF), которые соотносят результаты с распознанными цепями; Рекурсивные Сети Переходов (RTR), используемые большей частью для создания восходящих библиотек графов, т. е. повторного применения определенных графов внутри более сложных графов; Улучшенные Рекурсивные Сети Переходов (Enhanced Recursive Transition Networks), а именно RTR, которые содержат переменные – новшество, доступное уже в последних версиях INTEX, но значительно улучшенное в NooJ, позволяющем ввести лексические ограничения, которые могут отсылать к любой информации, содержащейся в словарях, для одной словарной формы слова или для набора словарных форм; и наконец, грамматики, независимые от контекста, или грамматики типа 2 (Context-Free Grammars), используемые прежде всего для морфологического описания. Разумеется, если не желателен графический интерфейс, есть возможность прибегнуть к регулярным выражениям.

NooJ предлагает также различные варианты дисплея словарей и таблиц лексики - грамматики. С позиций лексикографа, дисплей словаря NooJ в форме таблицы оказывается особенно удобным. Важно также, особенно из соображений касательно применения в преподавании, указать на то, что создание конкорданций и навигация между конкорданцией и полным текстом совершаются очень быстро, а параметры легко изменяемы. К тому же, программа включает морфологическую лабораторию, в которой приятно выполнять все виды морфологических упражнений на типологически разных языках.

В общем, эта мощная и быстрая программа для обработки естественного языка, опирающаяся на качественные и полные лингвистические ресурсы, в распоряжении научного сообщества на странице http://www.nooj4nlp.net/.

1. Blanco, Xavier; Silberztein, Max (2008): Proceedings of the 2007 International NooJ Conference, Cambridge Scholars Publishing.
2. Silberztein, Max (2004): “NooJ: A Cooperative, Object-Oriented Architecture for NLP”. in INTEX pour la linguistique et le traitement automatique des langues, Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté.
3. Silberztein, Max (2005): “NooJ’s Dictionaries” in Proceedings of LTC 2005, Poznan University.