Об аннотированных корпусах текстов

Л.Г.Митюшин
Лаборатория компьютерной лингвистики ИППИ РАН
mit@ippi.ac.msk.su

Я получил блаженное наследство -
Чужих певцов блуждающие сны...

О.Мандельштам

1. Введение

Повышение мощности компьютеров и успехи в области автоматической обработки текстов привели к появлению больших аннотированных корпусов - коллекций текстов, элементам которых приписана дополнительная информация. Примером может служить известный корпус Penn Treebank объемом около 4,5 млн слов, содержащий английские тексты вместе с их синтаксическим разбором [1].

В этой заметке описывается проект некоторой компьютерной системы, предназначенной для работы с аннотированным корпусом. Лингвистические данные, поддерживаемые системой, включают собственно аннотированный корпус и существующий независимо от него (но согласованный с ним) словарь. Тексты корпуса содержат информацию "ниже семантического уровня" - в частности, в них записываются морфологические данные о словоформах и синтаксические структуры предложений. В словаре сосредоточена семантическая информация. Система должна обслуживать три основные функции: ввод и коррекцию текстов корпуса, ввод и коррекцию статей словаря и обработку всевозможных запросов, относящихся к текстам корпуса.

Комбинация "аннотированный корпус плюс словарь" открывает интересные и разнообразные возможности для научных исследований и приложений. Создание подобных информационных массивов требует большого труда. Тем не менее представляется, что это один из тех путей, по которым должна идти лингвистика, чтобы приблизиться к естественнонаучным стандартам объективности.

2. Программа ввода текстов

Сейчас написана только программа для работы с текстами, реализующая первую из функций системы. В исходном виде тексты и все дополнительные данные представлены в форме, удобной для пользователя-лингвиста. Программа проверяет отсутствие противоречий во вводимой информации, преобразует ее в машинную форму, которая является одновременно компактной и удобной для компьютерной обработки, и записывает в специальную базу данных. Программа также выполняет обратную операцию, т.е. читает тексты из базы и преобразует их в исходную форму.

Программа позволяет вводить следующие данные о каждом слове текста:

1) имя лексемы;
2) часть речи;
3) морфологические характеристики;
4) данные об акцентуации;
пометы, отражающие особенности
данной словоформы;
6) комментарии.

Может также вводиться следующая информация об отношениях между словами:

7) синтаксические структуры в виде деревьев зависимостей со стрелками, помеченными именами синтаксических отношений;
8) данные о референтах местоимений;
9) при словах фраземы - ссылки на ее главное слово (при главном слове указывается имя фраземы);
10) ссылки на рифмы.

С помощью этой программы создан корпус стихов О.Мандельштама, в настоящее время (январь 1997 г.) включающий первые 100 стихотворений из двухтомника [2] (около 7000 слов). При вводе текстов используется вспомогательная программа морфологического и синтаксического анализа, которая строит "заготовки" аннотированных текстов с данными типа 1, 2, 3 и 7. Человек проверяет эти данные и вводит информацию остальных типов, после чего тексты записываются в корпус.

3. О словаре и обработке запросов

Программы для работы со словарем и запросами пока не написаны. Предполагается, что единицами словаря будут лексемы и фраземы, а также, возможно, терминологические и "статистически выделенные" сочетания (во многом похожие на фраземы). Корпус текстов связан со словарем через имена единиц. Считается, что система имеет общий словарь для всех текстов на данном языке, поддерживаемых с ее помощью.

Словарь должен содержать информацию, не присутствующую в явном виде в аннотированных текстах. Так, словарь не должен включать сведения о морфологических парадигмах, поскольку формы лексем (а также пометы о нестандартности форм, вариативности и т.п.) явным образом представлены в текстах. Словарь не должен содержать данных о

синтаксических свойствах лексем, так как эти свойства также "наблюдаемы" в аннотированных текстах.

В словарь записывается лингвистическая информация высших уровней. В нем фиксируются толкования единиц, семантические требования к их актантам, стилистические пометы. Должны также присутствовать разнообразные ссылки единиц друг на друга, отражающие, в частности, отношения лексической деривации, гипо- и гиперонимии, (квази)синонимии и антонимии, конверсии и т.п.

Заметим, что возможны и другие варианты распределения информации между текстами и словарем. Например, текстам могли бы приписываться не только синтаксические, но и семантические структуры. Предлагаемый вариант "минимизирует" информацию, помещаемую в корпус, при том, что сохраняется возможность решать интересные задачи.

Система обработки запросов должна выполнять команды вида "найти (или сосчитать) в текстах T ситуации, в которых имеет место явление F". Здесь T обозначает некоторое подмножество текстов корпуса, а F есть описание интересующих нас ситуаций в терминах информации, представленной в корпусе и словаре. Для задания T и F должен быть разработан достаточно мощный язык.

Пока такой системы нет, можно пользоваться программами, написанными ad hoc для каждого запроса. Это не так удобно, но эквивалентно с точки зрения получаемых результатов.

4. Возможные применения

4.1. Корпус как источник лингвистической информации

Укажем некоторые лингвистические задачи, которые можно решать, имея достаточно представительный аннотированный корпус описанного типа.

Если лингвиста интересует вопрос "возможны ли в языке ситуации F ?" или "насколько вероятны ситуации F ?", он сейчас чаще всего обращается к собственной языковой интуиции. Вместо этого он мог бы во многих случаях обращаться к корпусу.

Корпус может служить базой для настройки процедур автоматической обработки текстов, учитывающих вероятность тех или иных морфологических и синтаксических явлений. Примеры таких процедур: выработка гипотез о незнакомых словах, разрешение неоднозначностей, стохастический синтаксический анализ.

Корпус может отвечать на вопросы о лексической сочетаемости подобно тому, как это делает словарь BBI [3], и его можно использовать для тех же целей - например, человек, пишущий на неродном языке, может с его помощью находить идиоматичные способы выражения нужного смысла. Однако вопросы, которые можно "задавать корпусу", намного разнообразнее, чем в случае словаря. Кроме того, корпус позволяет получать статистическую информацию о комбинациях слов и, шире, о произвольных лексико-грамматических конфигурациях в текстах.

4.2. Исследование художественных текстов

Аннотированный корпус художественных текстов дает возможность анализировать количественными методами весьма широкий круг явлений. Например, в поэтических текстах можно исследовать ритм на различных уровнях, от фонетических признаков до семантических элементов (возможно, это приблизит нас к пониманию "музыки значений", о которой писал Ю.Н.Тынянов). Можно получать любые статистические данные - например, частоты стилистически значимых синтаксических явлений и конструкций.

Другая интересная возможность - автоматическое выявление интертекстуальных связей. Имеется в виду появление в разных текстах настолько близких друг к другу фрагментов, что вероятность их независимого возникновения слишком мала. "Близость" здесь следует понимать комплексно, с учетом факторов всех уровней. Разумеется, автоматически найденные случаи параллелизма должны рассматриваться только как гипотезы, подлежащие дальнейшему анализу.

Литература

[1] M.P.Marcus, B.Santorini, M.A.Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. // *Computational Linguistics*, 1993, Vol. 19, No. 2, 313-330.

[2] О.Мандельштам. *Сочинения*, тт. 1-2. "Художественная литература", М., 1990.

[3] M.Benson, E.Benson, R.Ilson. *The BBI Combinatory Dictionary of English*. John Benjamins, Amsterdam - Philadelphia, 1986.

On annotated text corpora

L.G.Mitjushin

Increasing power of computers and progress in automatic natural language processing made it possible to compile large annotated corpora, i.e. collections of texts with additional information assigned to their items. This note describes a design for a software system working with an annotated corpus.

Linguistic data within the system include a corpus per se and a lexicon. Texts of the corpus contain information "below the semantic level" (in particular, the morphological data on forms of the words and the syntactic structures of the sentences). The lexicon stores semantic information. The system should provide the following three types of operation: input and correction of texts, input and correction of lexicon entries, and processing of enquiries.

By now a program has been written only for the first of these functions. Using that program, an annotated corpus containing 100 poems by Osip Mandelstam has been compiled.