Извлечение информации о сочетаемости лексем из аннотированного корпуса текстов[1]

Дяченко Павел, Фролова Татьяна
Лаборатория компьютерной лингвистики ИППИ РАН
pavelvd@cl.iitp.ru, tfrolova@cl.iitp.ru

Аннотация

В настоящем докладе описывается извлечение из текстов информации о лексической сочетаемости слов, которая должна быть представлена в словарных статьях АОТ-систем для корректной обработки ими идиоматичных структур. Известно, что существует определенный набор синтаксических контекстов, в которых регулярно встречаются несвободные словосочетания. Таким образом, есть возможность получать информацию о несвободной лексической сочетаемости слов из синтаксически размеченного корпуса текстов. Для данного исследования используется морфологически и синтаксически размеченный корпус текстов «СинТагРус» объемом в 35 000 тысяч предложений, который был создан в Лаборатории компьютерной лингвистики ИППИ РАН. Из этого корпуса были автоматически выбраны определенные синтаксические контексты для пробного списка слов. Для пополнения и уточнения информации о сочетаемости в словарных статьях из результатов поиска были выбраны нестандартные словосочетания.

1. Лексическая сочетаемость слов и ее описание

В словарь системы машинного перевода, в частности, в словарь системы ЭТАП-3, должна быть помещена информация о способах анализа и перевода несвободных сочетаний слов. Например, в таких словосочетаниях как *бросать взгляд*, *накладывать ограничения*, *глубокий кризис* первое слово не может быть понято в своем основном значении и, возможно, должно переводиться на другие языки специальным образом. Поэтому в словарную статью одного из членов такого словосочетания вносится информация о том, как анализировать и переводить всё это словосочетание.

1.1. Лексические функции как способ описания несвободных словосочетаний

Было замечено, что существуют некоторые смыслы, которые выражаются идиоматично в разных языках и при разных словах. Для описания таких смыслов А.К.Жолковским и И.А.Мельчуком в [1,2] был предложен аппарат лексических функций. Например, прилагательное ГЛУБОКИЙ в словосочетании *глубокий кризис* в таком представлении является значением лексической функции MAGN. Лексическая функция MAGN ставит в соответствие слову X другое слово или словосочетание, обозначающее большую степень или интенсивность X-а (здесь и далее определения лексических функций приводятся по статье [3]). То же значение выражено нестандартным образом при других словах, ср. прилагательное БУРНЫЙ в словосочетании *бурная деятельность*, а также в других языках, ср. английское прилагательное SEVERE в словосочетании *severe crisis*. Было замечено также, что таких смыслов и, соответственно, таких лексических функций имеется лишь несколько десятков.

Значением лексической функции являются слова или словосочетания не только с определенным смыслом, но и выполняющие при данном слове X определенную синтаксическую функцию. Так, например, значением лексической функции MAGN для существительных являются прилагательные (ср. *глубокий* кризис) выполняющие при X-е функцию синтаксического определения.

[1] Работа частично поддержана средствами грантов РФФИ № 08-06-00344 и РФФИ № 07-06-00339.

1.2. Представление информации о лексической сочетаемости в комбинаторном словаре системы ЭТАП-3

В комбинаторном словаре системы ЭТАП-3 (более подробно о системе см. [4-6]) информация о способе анализа и перевода лексикофункциональных словосочетаний должна быть помещена в словарную статью слова X в специальной зоне лексических функций. Так, в словарной статье слова КРИЗИС в зоне лексических функций имеется запись «MAGN: *глубокий*». В комбинаторном словаре другого языка переводной эквивалент русского слова также снабжен информацией о способе выражения лексической функции MAGN, например, в словарной статье слова CRISIS английского комбинаторного словаря имеется запись «MAGN:*severe*». Это обеспечивает возможность адекватного перевода данной структуры.

Информация о нестандартной сочетаемости лексем, не поддающейся классификации в рамках системы лексических функций, описывается специальными правилами синтаксического анализа и перевода в словарной статье одного из членов такого словосочетания.

1.3. Источники сведений о лексической сочетаемости

В словарь системы ЭТАП-3 информация о лексической сочетаемости, в частности, информация о лексических функциях поступает либо непосредственно от составителя словаря (лингвиста и/или носителя данного языка) в результате интроспекции, либо из словарей данного языка: словарей сочетаемости, а также толковых и двуязычных словарей.

В последнее время, благодаря появлению корпусов текстов с разной глубиной разметки, стало возможным использовать свойство лексических функций выступать в стандартных синтаксических контекстах и получать информацию о нестандартной сочетаемости лексем из текстов. Так, например, для получения данных о лексической функции MAGN и других адъективных лексических функциях для данного слова можно осуществлять поиск слов и словосочетаний, зависящих от данного по определительному отношению в синтаксически размеченном корпусе.

2. Пробный поиск несвободных словосочетаний в корпусе текстов

В рамках данного исследования был осуществлен поиск данных о значениях адъективных лексических функций для ряда существительных и о тех нестандартных коллокатах этих существительных, которые имеют схожее с лексическими функциями синтаксическое оформление, на материале морфологически и синтаксически размеченного корпуса текстов «СинТагРус» объемом в 35 000 тысяч предложений, который был создан в Лаборатории компьютерной лингвистики ИППИ РАН. Использовалась полная версия корпуса по состоянию на июнь 2008 года.

Выбор контекстов с определенными синтаксическими условиями и сортировка этих контекстов по встречающимся в них прилагательным осуществлялся автоматически.

Алгоритм поиска представляет собой последовательный просмотр всех предложений размеченного корпуса. Для каждого предложения проверялось выполнение контекстного условия, которое на данном этапе исследования было сформулировано заранее и описано непосредственно в коде программы. В то же время, были выделены настраиваемые параметры поиска – например, объем просматриваемого корпуса, тип синтаксической связи между узлами контекста, а также морфологические характеристики интересующих нас коллокатов. Отобранные программой поиска контексты автоматически сортировались по частотности и сохранялись для последующей обработки.

Выбор нестандартных коллокатов, данные о которых должны быть помещены в словарь, осуществлялся вручную.

2.1. Условия поиска

Основные результаты для адъективных лексических функций и подобных им сочетаний получены в результате автоматического анализа определительных контекстов. То есть по заданному существительному находились прилагательные, зависящие от заданного существительного по определительному отношению, такие как прилагательное ГЛУБОКИЙ в словосочетании *глубокий кризис*. Ограничения на часть речи синтаксического определения не вводились, однако на пробном списке существительных были найдены и отобраны как нестандартные только словосочетания, в которых определительным отношением с заданными существительными были связаны прилагательные. Определения, относящиеся к другим частям речи, составили с заданными существительными свободные словосочетания, которые нет необходимости специальным образом описывать в словаре.

Для получения более полного результата выполнялся также поиск с более сложными синтаксическими условиями. Например, поиск прилагательных, которые связаны с синтаксическим определением заданного слова сочинительным отношением, таких как прилагательное ГЛУБОКИЙ в словосочетании *продолжительный и глубокий кризис*. А также прилагательных, выступающих в предикативной позиции в именных предложениях и предложениях с глаголом-связкой, таких как прилагательное ГЛУБОКИЙ в предложениях *кризис глубок* или *кризис был глубоким*.

Таким образом, всего для адъективных коллокатов было рассмотрено три типа контекстов. Такое расширение параметров поиска позволило получить очень небольшое количество дополнительных результатов (всего получено 222 нестандартных сочетания с прилагательными, из них лишь одно – в результате дополнительного поиска по контекстам с сочинительными и предикативными синтаксическими отношениями).

2.2. Интерпретация результатов

Из полученных 222 словосочетаний около четверти (53) могут быть описаны как сочетание существительного со значением лексической функции MAGN (определение см. выше), например, словосочетание *серьезное обвинение*. Двенадцать прилагательных являются значениями лексической функции BON (выражает стандартную положительную оценку X-а) от соответствующих существительных, например, словосочетание *теплый прием*. В одиннадцати случаях полученные прилагательные можно описать как значение лексической функции ANTIMAGN (обозначает небольшую степень X-а), например, в словосочетании *небольшое количество*. В трех случаях, в словосочетаниях *ошибочный подход*, *плохая работа*, *черное дело*, прилагательные являются значениями лексической функции ANTIBON (выражает стандартную отрицательную оценку X-а), в двух случаях, в словосочетаниях *прямое назначение*, *трезвый расчет* – значениями лексической функции VER (имеет значение свойства, которое нормально ожидается или требуется от X-а), в двух словосочетаниях *бесплодный поиск*, *неосуществленный проект*, – значениями лексической функции ANTIVER (имеет значение свойства, противоположного тому, которое нормально требуется или ожидается от X-а). Более половины словосочетаний (139) не могут быть описаны в терминах лексических функций и должны обслуживаться в словаре особыми правилами перевода, например, словосочетание *лесное хозяйство*.

3. Оценка результатов экспериментального поиска и их практическое применение

Полученные результаты были сопоставлены с теми сведениями о сочетаемости слов, которые приводятся в русском электронном словаре ABBYY Lingvo 12 и в Большом русско-английском словаре под ред. А. И. Смирницкого, представленном в электронном словаре МультиЛекс. В то же время, результаты эксперимента были применены для улучшения качества автоматического перевода в системе ЭТАП-3.

3.1. Сравнение с лексикографическими источниками

Из 222 несвободных словосочетаний, найденных в результате пробного поиска по корпусу, только 119 имеется хотя бы в одном из двух упомянутых выше словарей, а остальные 103 словосочетания не представлены ни в одном из них.

Таким образом, для описания лексической сочетаемости в словарных статьях текстовые данные представляют собой существенное дополнение к лексикографическим источникам.

3.2. Практическое применение результатов эксперимента в системе ЭТАП-3

Автоматический перевод выделенных словосочетаний на английский язык при помощи системы ЭТАП-3 дал следующие результаты: для 120 словосочетаний из выделенных 222 был получен удовлетворительный перевод, остальные 102 словосочетания были переведены неверно. После внесения информации об этих 102 словосочетаниях в русский комбинаторный словарь системы ЭТАП-3 был получен правильный перевод.

4. Перспективы метода

Поиск коллокатов слов на материале синтаксически размеченного корпуса предполагается развивать как путем расширения списка слов-аргументов лексических функций, так и путем добавления к условиям поиска других синтаксических контекстов, характерных для глагольных, субстантивных и адвербиальных

лексических функций. Ниже приведены примеры дельнейшего применения метода поиска по контекстам.

4.1. Поиск по значениям лексических функций

Поиск по текстам можно использовать и другим образом. Если для составления лексикографических портретов слов актуален поиск различных лексикофункциональных контекстов данного слова, то для систематического пополнения словаря поиск может быть «перевернут».

Несмотря на то, что выбор данного значения лексической функции для данного слова не может быть с точностью предсказан, имеется тенденция к семантической мотивированности этого выбора, как это было показано в [7-9]. То есть определенное значение лексической функции может быть характерным для целого класса слов-аргументов, обладающих общими семантическими признаками. Так, в 83 лексикофункциональных словосочетаниях, найденных в результате пробного поиска, описанного выше, встречается только 51 прилагательное.

Таким образом, поиск синтаксически схожих несвободных словосочетаний можно осуществлять не только по словам-аргументам, но и по значениям лексических функций, повернув стрелки в условиях поиска. Например, по прилагательному ОЖЕСТОЧЕННЫЙ, являющемуся значением лексической функции MAGN от существительного БОРЬБА, можно найти все существительные, от которых данное прилагательное зависит по определительному отношению. Такой поиск даёт ещё четыре лексикофункциональных словосочетания, информация о которых должна быть помещена в словарные статьи существительных СТОЛКНОВЕНИЕ, СОПРОТИВЛЕНИЕ, БОЙ и КОНКУРЕНЦИЯ.

При этом поиск по корпусу имеет некоторое преимущество по сравнению с извлечением информации из словарей. При поиске по корпусу словосочетания будут найдены независимо от того, заданы ли условия поиска по аргументам или по значениям лексических функций.

В то же время, например, при составлении лексикографического портрета слова по материалам словарей часть приведенных в словарях словосочетаний не будет найдена, поскольку эти словосочетания описываются в словарных статьях значений лексических функций, а не аргументов (ср. описание сложностей, возникающих в связи с таким свойством словарей в процессе обучения языку в [10]). Так, например, из 222 словосочетаний, представленных в словарях и одновременно являющихся результатами пробного поиска (ср. пункт 3 выше), только 36 находятся в словарных статьях соответствующих существительных, остальные 83 помещены в словарные статьи прилагательных, поэтому узнать об их наличии из словарной статьи существительного невозможно.

4.3. Поиск в морфологически размеченном корпусе

Дальнейшей перспективой развития метода является поиск в корпусе с морфологической разметкой. Поиск по Национальному корпусу должен дать значительно больше результатов, чем поиск по синтаксически размеченному корпусу, из-за существенно большего объема этого корпуса. В то же время, если задавать условия поиска только в терминах возможных морфологических признаков и возможного относительного расположения слов в предложении, уровень «шума» будет достаточно высок. Для исключения некоторого количества лишних контекстов из результатов поиска по Национальному корпусу кажется оправданным наложить дополнительные ограничения на относительное расположение и морфологические характеристики слов в словосочетаниях. Результаты поиска по синтаксически размеченному корпусу позволяют сформулировать предположения относительно содержания этих ограничений.

Так, например, выше было замечено, что, хотя значением лексической функции MAGN в принципе могут являться не только прилагательные, в итоговый список нестандартных определений-коллокатов для пробного списка существительных попали только они. С другой стороны, из 471 вхождения 222 несвободных словосочетаний только в 42 между прилагательным и существительным встречались другие слова, и только в 6 прилагательное следовало за существительным. Таким образом, ограничив поиск только прилагательными и только непосредственно предшествующими существительному, можно существенно снизить уровень «шума» в результатах поиска, лишь в достаточно небольшой степени пожертвовав полнотой материала.

[1] А.К. Жолковский, И.А. Мельчук. «О семантическом синтезе», *Проблемы кибернетики. Вып. 19*. М.: Наука, 1967, с. 177-238.

[2] И.А. Мельчук. *Опыт теории лингвистических моделей «Смысл – Текст»*. М., Наука, 1974.

[3] Ю.Д. Апресян, П.В. Дяченко, А.В. Лазурский, Л.Л.Цинман. «О компьютерном учебнике лексики русского языка», *Русский язык в научном освещении*. №2(14), 2007, с. 48-112.

[4] 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*. М: Наука, 1989.

[5] 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*. М: Наука, 1992.

[6] Ju.D. Apresian, I.M. Boguslavsky, L.L. Iomdin, A.V.Lazursky, V.Z. Sannikov, V.G. Sizov, L.L. Tsinman. «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, 2003, p. 279-288.

[7] Ю.Д. Апресян. «О семантической непустоте и мотивированности глагольных лексических функций», *Вопросы Языкознания № 4*, 2004, с. 3-18.

[8] Ю.Д. Апресян. «Лексические функции на службе компьютерной лингвистики» *Прикладна лінгвістика та лінгвістичні технології MEGALING-2006. Збірник наукових праць*. Київ, 2007, с. 7-20.

[9] Ю.Д. Апресян. «О семантической мотивированности лексических функций-коллокатов», *Вопросы Языкознания № 5*, 2008, (в печати).

[10] *Margarita Alonso Ramos*. «Glosas para las colocaciones en el Diccionario de Colocaciones del Español». *Diccionario y Fraseología*. *ed. by M. Alonso Ramos*. Coruña: Universidade da Coruña, 2006, p. 59-88.