Словари сочетаемости слов: какой метод составления лучше?[1]

А. Ф. Гельбух, Г. О. Сидоров, Э. Эрнандес-Рубио
Лаборатория обработки естественного языка,
Центр Вычислительных исследований (CIC),
Национальный Политехнический Институт (IPN),
г. Мехико, Мексика
gelbukh@gelbukh.com, sidorov@cic.ipn.mx

М. В. Чубукова
Независимый исследователь, г. Москва, Россия

В докладе рассматривается проблема автоматического построения словарей сочетаемости. Проводится сравнение ручного и автоматического подходов к пополнению словарей сочетаемости. Представлен синтаксический анализатор, основанный на контекстно-свободной грамматике для испанского языка, использованные в эксперименте. Кратко представлена среда разработки подобного типа КС-грамматик. Показано, что синтаксический анализ является желательным этапом при автоматическом пополнении словарей сочетаемости, для чего проводится количественное сравнение метода пополнения словаря сочетаемости, основанного на синтаксическом анализе, и метода, основанного на биграммах. Для количественной оценки был проведен эксперимент, в котором для случайно выбранного текста на испанском языке был проделан автоматический синтаксический анализ и извлечены словосочетания, которые должны быть добавлены в словарь сочетаемости. Затем на том же тексте выделение словосочетаний было проведено вручную и с использованием метода биграмм. На основе сравнения с ручной разметкой были подсчитаны точность и полнота метода с синтаксическим анализом и биграммного метода, которые существенно лучше у метода, использующего синтаксический анализ.

–... А диван?
– Безотказен. Конструкции Льва Бен Бецалеля. Бен Бецалель собирал и отлаживал его 300 лет.
– Вот. Учитесь. Старик, а всё делал сам.

[1] Работа выполнена при частичной поддержке правительства Мексики (Конацит, СНИ) и Национального политехнического института (CGPI, COFAA). Мы благодарим И.А. Большакова за полезные дискуссии. Имена авторов даны в алфавитном порядке. Work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (CGPI, COFAA). We thank Prof. Igor A. Bolshakov for useful discussions. Authors' names are given in alphabetic order.

А. и Б. Стругацкие «Понедельник начинается в субботу»

Введение

Нужны ресурсы. Примерно так говорят разработчики систем, использующих лингвистические знания. И чем больше, тем лучше. Под ресурсами понимаются лингвистические базы данных, словари разных типов, корпуса, формальные лингвистические модели (например, формальные грамматики), в общем, все, что содержит хоть какую-то информацию о словах или может быть использовано для получения такой информации. Почему это так важно? Именно информация о словах позволяет строить языковые представления более высоких уровней. На каждом уровне языка слово имеет свою информацию – фонетическую, морфологическую, синтаксическую, семантическую, прагматическую. Два первых и последний тип очевидно находятся вне рамок данной статьи. Под семантической информацией можно иметь в виду информацию, которая содержится, скажем, в толковых и переводных словарях в виде определений слов или в виде набора семантических признаков. Интересно, что переводные словари тоже являются важным ресурсом, потому что отражают неконгруэнтное представление мира словами в разных языках, что может использоваться при автоматической обработке, см., например, (Гельбух, 1997). Синтаксическая информация отражает, прежде всего, сочетаемостные характеристики слов – словари словосочетаний, модели управления (что является обобщенным словарем словосочетаний, например, сочетаемость с существительными в дательном падеже), синтаксические роли, соответствие семантических и синтаксических валентностей в стиле ТКС, словари идиом и лексических функций (т.е. специализированные словари словосочетаний). В дальнейшем будем рассматривать только словари словосочетаний в прямом смысле, то есть те, которые содержат исключительно сочетания слов. На их основе можно строить другие типы словарей, содержащих синтаксическую информацию.

Теперь обсудим вопрос, в каких приложениях можно использовать синтаксическую информацию. Эта информация полезна в большинстве приложений, связанных с обработкой естественного языка, потому что дополнительная априорная информация о сочетаемости слов позволяет разрешать неоднозначность, а именно неоднозначность представляет наибольшие проблемы при автоматической обработке. Рассмотрим несколько примеров. Сначала классический случай референциальной неоднозначности.

Иван взял хлеб со стола и уронил его.

В этом случае местоимение *его* относится очень вероятно к *хлебу*, и маловероятно, что к *столу*. В отличии от другого предложения

Иван взял хлеб со стола и протер его.

где имеется обратная ситуация. Люди догадываются о правильной референции, используя знания о мире, где не очень часто встречаются ситуации *уронить стол* или *протереть хлеб.* Хотя такое возможно, но очень маловероятно. Компьютер не имеет подобных знаний, но их можно промоделировать посредством словаря словосочетаний, где существуют словосочетания *протереть стол* и *уронить хлеб*. Как вариант, можно иметь словосочетания *протереть мебель* и *уронить еду*, и дополнительно применять правила вывода на основе знаний на уровне тезауруса, что *стол* это *мебель*, а *хлеб* это *еда*.

Другой пример возможного применения словаря словосочетаний это машинный перевод. В идеале, имея две очень большие базы словосочетаний и соотношение между ними можно решать многие проблемы перевода. Но даже не очень большой словарь может помочь в некоторых случаях, например, при переводе лексических функций (Mel’chuk, 1995) – скажем,

to pay attention (букв., *платить внимание*) должно быть переведено на русский как *обратить внимание*, а на испанский как *prestar atención* (букв., *одолжить внимание*) и т.п.
Еще один пример использования словаря словосочетаний – разрешение синтаксической неоднозначности. Например, *рассмотреть вопрос об инвестициях в России*.
Неоднозначность состоит в том, какое слово является хозяином для *в России*: это может быть *вопрос* или это может быть *инвестиции*. Если в нашей базе данных будет соответствующее словосочетание, то появляются основания разрешить эту неоднозначность.
Далее в статье сравниваются различные методы построения словарей словосочетаний, кратко представляется синтаксический анализатор и описывается среда разработки грамматики, затем приводятся данные эксперимента, количественно оценивающего подход, основанный на биграммах и подход, основанный на синтаксическом анализе.

Методы построения словарей словосочетаний

Итак, проблема состоит в том, как получать построить (пополнять) словарь словосочетаний. Идеальный вариант – много высококвалифицированных лингвистов пишут словарь, который содержит всю полезную синтаксическую информацию. См. эпиграф. К сожалению, полностью идеальный вариант не достижим по вненаучным соображениям. Тем не менее, какая-то часть идеального варианта существует в виде готовых созданных вручную словарей словосочетаний, например, для русского языка это система КроссЛексика (Bolshakov and Gelbukh, 2001, 2002). Система содержит более миллиона словосочетаний и не очень сложный набор семантических связей. В систему встроена возможность вывода на основе тезаурусных отношений. Для английского языка также существуют подобные ресурсы, хотя и не такие большие, например, словарь Oxford (Oxford, 2003) содержит около 170,000 единиц, словарь Collins (Bank of English, 2003) около 140,000 единиц, без возможностей вывода.
Есть и автоматические методы извлечения словосочетаний из текста (см. работы 1, 5, 9, 10, 13, 14). Но все эти автоматические методы основаны на взаимной информации слов частотах из совместной встречаемости в текстах. Это значит, что будут находиться только словосочетания с достаточно большой частотой. А даже в очень больших корпусах текстов и большинство-то слов не имеет высоких частот в соответствии с законом Ципфа, тем более гораздо меньшие частоты будут иметь сочетания слов. Например, одной из широко известных работ этого типа является статья (Smadja, 1993), где описывается система *Xtract*. В этой системе предусмотрено три стадии анализа, и на последней стадии даже применяется частичный синтаксический анализ для отбора синтаксически связанных словосочетаний. Тем не менее, как уже было сказано, отбираются только высокочастотные словосочетания, и, тем самым, подавляющее большинство словосочетаний просто не учитывается.
Еще один возможный метод получения словосочетаний состоит в использовании биграмм, то есть, стоящие рядом слова рассматриваются как словосочетания. В этом случае нет зависимости от размера корпуса и частот слов. Однако заранее можно предсказать, что такой метод будет иметь низкую точность, потому что слова, стоящие рядом в предложении, не всегда являются словосочетаниями.
В данной статье предлагается использовать другой автоматический метод извлечения словосочетаний, основанный на автоматическом синтаксическом анализе. На основе полученного дерева зависимостей можно извлечь правильные словосочетания независимо от взаимного положения слов в предложении. Дополнительно необходимо особым образом обрабатывать:

- обороты с сочинительной связью, потому что обе части должны быть добавлены в словарь с одним и тем же главным словом;

- предлоги, потому что предлоги не характеризуют лексическую сочетаемость, а выражают грамматические отношения, которые также могут выражаться, скажем, падежами. Тем не менее, информация о том, какой именно предлог используется, ценна. Так что в этом случае хранится трехчленное словосочетание – главное слово + предлог + зависимое слово.

При этом методе можно добавлять в словарь дополнительную полезную информацию. Скажем, в случае управления важно хранить грамматическую информацию о зависимом слове, потому что она может меняться в зависимости от главного слова. При согласовании такая информация несущественна. В случае испанского языка для существительных хранится информация о числе. Для русского языка должна еще хранится информация о падеже. Для глагола хранится обобщенная информация о его форме – инфинитив, деепричастие или причастие без дальнейшей детализации (напомним, что глагол в личной форме не может быть зависимым словом).

Кроме того, в словаре хранится тип синтаксического отношения. В данной версии грамматики используются следующие отношения: *dobj* (прямое дополнение), *subj* (субъект), *obj* (косвенное дополнение), *det* (модификатор – артикль или местоимение), *adver* (наречие), *cir* (обстоятельство), *prep* (предлог), *mod* (модификатор, отличный от артикля или местоимения), *subord* (подчинение), *coord* (сочинение).

Дополнительные возможности при использовании подхода с синтаксическим анализом состоят в том, что можно фильтровать словосочетания, используя типы синтаксических отношений и морфологическую информацию. Очевидно, что словосочетания, скажем, с союзом или местоимением нет смысла добавлять в словарь. Кроме того, можно не учитывать, например, обстоятельственное отношение, или, в зависимости от потребностей пользователя, какой-нибудь другой тип отношений.

Сравнивая автоматический и ручной подходы можно отметить следующее. Очевидное преимущество ручного подхода к составлению словарей сочетаемости – его высокое качество, тем не менее также ясны и его недостатки – субъективность (разные люди принимают разные решения в сходных ситуациях), непоследовательность (на человека могут влиять различные факторы и он может принимать разные решения), неполнота, низкая скорость, высокая стоимость. Преимущества автоматического подхода – полнота (зависит от обрабатываемого корпуса), большая скорость, последовательность, объективность, низкая стоимость. Но есть и весьма существенный недостаток – качество построения словаря словосочетаний зависит от качества синтаксического анализа, которое пока не настолько высоко, как того хотелось бы.

Синтаксический анализатор

Синтаксический анализатор состоит из языково-независимой программной оболочки и наборов данных для конкретного языка, включающих следующие массивы:

- Таблица базы данных морфологического анализатора, состоящая из трех колонок: словоформа – лемма (нормализованная форма) – морфосинтаксическая информация о словоформе,
- Грамматика в разработанном авторами формализме, основанном на формализме контекстно-свободных грамматик, но с элементами унификации.

Среда позволяет видеть результаты морфологического разбора и варианты синтаксического разбора в различных представлениях (см. Рис. 1). При синтаксическом анализе дается возможность проследить детали внутреннего поведения анализатора, то есть локализовать те места, в которых, в случае необходимости, нужно вносить изменения в грамматику. Более подробно среда описана в (Gelbukh *et al.*, 2002).
Ядром среды является модуль синтаксического анализа (парсер), который использует контекстно-свободную грамматику с элементами унификации. Парсер реализован на основе алгоритма chart parser. Дополнительно, парсер присваивает вариантам веса в соответствии со статистической информацией, полученной для испанского языка на основе обсчета корпуса LEXESP. Для получения весов были автоматически проанализированы модели управления глаголов, прилагательных и некоторых существительных (Gelbukh *et al.*, 1998). Эти веса позволяют упорядочить варианты разбора от более вероятных к менее вероятным.

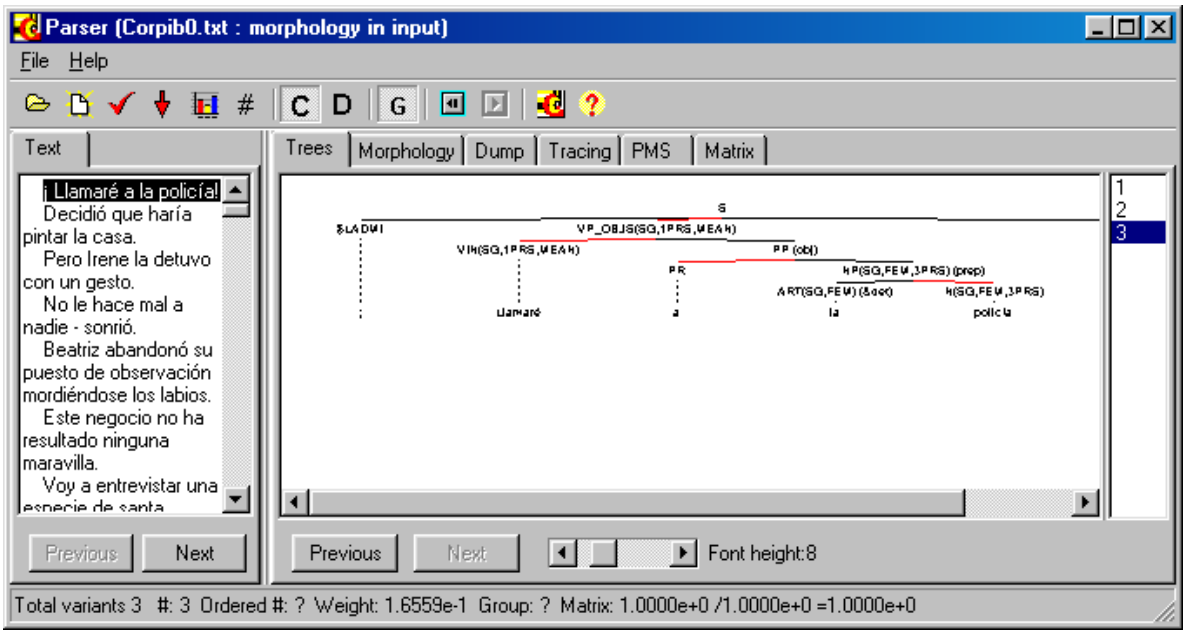

Рис. 1. Значения весов вариантов.

Эксперимент

Эксперимент проводился на случайно выбранном испанском тексте. Всего в тексте было разобрано 60 предложений, в которых было 741 слово, в среднем 12.4 слова на предложение. Для получения возможности сравнения автоматических методов была проведена ручная разметка синтаксических зависимостей.
Рассмотрим пример. *Conocía todos los recovecos del río y sus misterios. (Он знал все излучины реки и ее тайны.)* Парсер возвращает следующее дерево зависимостей, в распечатке которого зависимости помечены не стрелками, а глубиной вложения, то есть, слова, зависящие от одного и того же слова, находятся на одном и том же вертикальном уровне (с тем же отступом от левого края). Например, V(SG, 1PRS, MEAN) [*conocía* (*знал*)] является вершиной предложения, и от него зависят союз CONJ_C [*y* (и)] и точка $PERIOD. От союза CONJ_C зависят слова, находящиеся на следующем уровне, N(PL, MASC) [*recovecos* (*излучины*)] и N(PL,MASC) [*misterios* (*майны*)], и т.д. На каждой строчке присутствует морфологическая категория, использованная в грамматике, форма слова и лемма. В фигурных скобках стоит имя синтаксического отношения.
1 **V(SG,1PRS,MEAN)** // *Conocía : conocer* (*знал : знать*)

2 ...**CONJ_C** {dobj} // *y : y* (*и : и*)
3 .....**N(PL,MASC)** // *recovecos : recoveco* (*излучины : излучина*)
4 .....**PR** {prep} // *del : del* (предлог соответствующий родительному падежу)
5 .....**N(SG,MASC)** {prep} // *río : río* (*реки : река*)
6 .....**ART(PL,MASC)** {det} // *los : el* (*определенный артикль*)
7 .....**#*$$todo#** // *todos : todo* (*все : все*)
8 .....**N(PL,MASC)** // *misterios : misterio* (*тайны : тайна*)
9 .....**DET(PL,MASC)** {det} // *sus : su* (*ее : он*)
10....**$PERIOD** . // . : .

Были найдены следующие словосочетания, не считая сочетаний со служебными частями речи, где первое слово является главным, а последнее – зависимым: *conocer* (dobj) *recoveco* {Pl} (*знать + излучину*), *conocer* (dobj) *misterio* {Pl} (*знать + тайну*), *recoveco* (prep) [*del*] *río* {Sg} (*излучина + река* (+ Род.п., переданный предлогом в испанском)).

Как можно видеть, отношение *dobj* соответствует синтагме с сочинительным союзом, и поэтому продвигается до хозяина этого союза, у которого зависимыми словами становятся: *recovecos* (*излучины*) и *misterio* (*тайна*). Предлог *del* является частью третьего словосочетания. Словосочетания с артиклями и местоимениями (*el, todo, su*) были отфильтрованы, хотя они и находятся алгоритмом.

Для лучшего сравнения методов были внесены некоторые улучшения в метод, основанный на биграммах – была добавлена возможность пропускать артикли и учитывать предлоги в качестве третьего члена словосочетаний. Всего в тексте было 153 артикля и предлога, т.е. количество слов, участвующих в биграммном методе 741 – 152 = 588.

Были получены следующие результаты. Количество правильных словосочетаний, размеченных вручную, составляет 208. Из них, методом с синтаксическим анализом были найдены 148, а биграммный метод нашел 111. Но метод с синтаксическим анализом ошибся в 63 случаях (то есть нашел 63 неправильных словосочетания), а биграммный метод выдает 588 * 2 – 1 = 1175 словосочетаний, большинство из которых неправильные.

Это дает следующие значения параметров точности и полноты (precision и recall). Напомним, что точность это отношение правильных полученных ко всем полученным, а полнота – отношение правильных полученных ко всем правильным (неважно, получены они или нет). Это дает такие цифры. Метод с синтаксическим анализом имеет точность 148 / (148 + 63) = 0.70 и полноту 148 / 208 = 0.71; точность же биграммного метода 111/1175 = 0.09, а его полнота равна 111 / 208 = 0.53. Как видно, точность метода с синтаксическим анализом почти на порядок лучше, а полнота несколько лучше (примерно на 20 процентов).

Выводы

В статье было показано, что словари словосочетаний являются полезным лингвистическим ресурсом. Были рассмотрены различные способы составления словарей словосочетаний – ручной, автоматический, основанный на синтаксическом анализе, и автоматический, основанный на биграммах. Кратко была описана среда для разработки контекстно-свободных грамматик для естественных языков, одна такая грамматика для испанского языка использовалась в эксперименте.

Было приведено описание эксперимента по извлечению словосочетаний из текста на испанском языке, в результате которого были получены количественные оценки, показывающие преимущество метода с синтаксическим анализом над биграммным методом. Для оценки использовались данные ручной разметки и результаты синтаксического разбора испанского текста.

Литература

1. Baddorf, D. S. and M. W. Evens. Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In: *Proc. of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS'98)*, Dayton, USA, 1998.
2. Bank of English. Collins. http://titania.cobuild.collins.co.uk/boe_info.html
3. Bolshakov, I. A., A. Gelbukh. A Very Large Database of Collocations and Semantic Links. In: Mokrane et al. (Eds.) Natural Language Processing and Information Systems, 5th International Conference on Natural Language Applications to Information Systems NLDB-2000, Versailles, France, June 2000. Lecture Notes in Computer Science No. 1959, Springer Verlag, 2001, p. 103-114.
4. Bolshakov, I. A., A. Gelbukh. Word Combinations as an Important Part of Modern Electronic Dictionaries. *Revista SEPLN (Sociedad Español para el Procesamiento del Lenguaje Natural)*, No. 29, septiembre 2002, p. 47-54.
5. Dagan, I., L. Lee, and F. Pereira. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1), 1999.
6. Gelbukh, A. *Using a semantic network for lexical and syntactical disambiguation*. Proc. CIC-97, Nuevas Aplicaciones e Innovaciones Tecnológicas en Computación, Simposium Internacional de Computación, 1997, CIC, IPN, Mexico City, Mexico, pp. 352–366.
7. Gelbukh, A., G. Sidorov, S. Galicia Haro, I. Bolshakov. Environment for Development of a Natural Language Syntactic Analyzer. In: Acta Academia 2002, Moldova, 2002, pp.206-213.
8. Gelbukh, A., I. Bolshakov, S. Galicia Haro. Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts. *Int. Conf. on Automatic Learning and Discovery*, Pittsburgh, USA, June 11 - 13, pp. 261 - 267, 1998.
9. Kim, S., J. Yoon, and M. Song. Automatic extraction of collocations from Korean text. *Computers and the Humanities* 35 (3): 273-297, August 2001, Kluwer Academic Publishers.
10. Kita, K., Y. Kato, T. Omoto, and Y. Yano. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, 1(1):21-33, 1994.
11. Mel'čuk, I.Phrasemes in language and phraseology in linguistics. In: *Idioms: structural and psychological perspective*, 1995, pp. 167-232.
12. Oxford collocation dictionary, Oxford, 2003.
13. Smadja, F. Retrieving collocations from texts: Xtract. *Computational linguistics*, 19 (1):143-177, March 1993.
14. Yu, J., Zh. Jin, and Zh. Wen. Automatic extraction of collocations. 2003.