Сравнение методов автоматического выделения глагольно-именных словосочетаний

С.С. Кощеева

Санкт-Петербургский государственный университет
happybrightsv@yandex.ru

Аннотация

В работе описывается ряд экспериментов, проведённых с целью сравнить методы автоматического выделения устойчивых глагольно-именных словосочетаний (коллокаций). Делаются выводы о влиянии ширины диапазона и учёта частеречной принадлежности коллоката на качество выделения словосочетаний.

1. Введение

Корпусно-ориентированные исследования коллокаций представляют собой один из наиболее современных и развивающихся подходов к изучению явлений устойчивости и сочетаемости лексических единиц. Выделение устойчивых словосочетаний находит применение во многих областях, среди которых наиболее важными являются семантические и лексикографические исследования (в том числе, создание словарей и грамматик нового типа), преподавание языков и перевод, машинный перевод, автоматический анализ и снятие неоднозначности.

Использование статистического подхода для поиска и выделения устойчивых словосочетаний получило широкое распространение в корпусной лингвистике. Наиболее простым способом выявления коллокаций в тексте является составление частотных списков слов, оказавшихся слева или справа от ключевого в пределах заданного диапазона. Этот диапазон обычно составляет + – 5 слов.

В настоящее время широкое распространение получили статистические меры ассоциации (log-likelihood, MI, t-score и др.), которые основаны на формулах, использующих частоту совместного появления слов в коллокации, частоты каждого компонента словосочетания, объем корпуса и др. (см., напр., [6]). При этом эти частоты могут подсчитываться также в пределах некоторого диапазона.

В качестве меры ассоциации и ранжирования при выделении словосочетаний была выбрана мера Mutual Information (MI) [4], которую можно определить как коэффициент силы синтагматической связанности [5]. Мера MI (Mutual information) позволяет определить, насколько значимой является встречаемость двух слов на основе сравнения частоты совместного появления двух слов и произведением частот их независимого появления в тексте:

$MI = \log_2 \frac{f(n,c)*N}{f(n)*f(c)}$, где

n – node (ключевое слово);

c – collocate (коллокат);

f(n,c) – частота встречаемости ключевого слова n в паре с коллокатом c;

f(n), f(c) – абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте);

N – общее число словоформ в корпусе (тексте).

Однако необходимо помнить о том, что слова находятся в определённых отношениях (синтаксических, семантических) и не появляются в тексте (корпусе текстов) совершенно случайно. Следовательно, выделение коллокаций требует использования не только статистических методов, но и учёта морфологических и синтаксических свойства исследуемых единиц.

Целью данной работы является изучение и сравнение методов автоматического выделения глагольно-именных словосочетаний. Предметом данного исследования являются глагольно-именные словосочетания типа глагол + существительное (в винительном падеже без предлога). Инструментом для выделения глагольно-именных коллокаций послужило программное средство IntelliText, разработанное Центром переводческих исследований (Centre for Translation Studies) в университете г. Лидса (http://corpus.leeds.ac.uk/it/). Особенности данной системы – богатые лингвистические возможности и наличие репрезентативных корпусов, в том числе морфологически размеченных корпусов русского языка. В качестве материала для нашего исследования был использован русскоязычный корпус текстов RNC2010-MOCKY, версия Национального корпуса русского языка (НКРЯ) 2010 г. объёмом 116 млн слов.

2. Методика исследования

Эксперименты по выделению глагольно-именных словосочетаний проводились для глаголов *выполнять, нарушать, принимать*. На основе Словаря сочетаемости слов русского языка (под ред. П.Н. Денисова и В.В. Морковкина) [3], Словаря устойчивых глагольно-именных словосочетаний рус-

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

ского языка (В.М. Дерибас) [1] и Малого академического словаря (под ред. А.П. Евгеньевой) [2] был составлен список устойчивых глагольно-именных словосочетаний (для глаголов *выполнять, нарушать, принимать*), с которыми сравнивались коллокации, выделенные системой IntelliText. Ниже представлен список глагольно-именных словосочетаний для глагола *выполнять*:

1. выполнять директивы
2. выполнять долг
3. выполнять желание
4. выполнять задание
5. выполнять задачу
6. выполнять заказ
7. выполнять заявку
8. выполнять инструкцию
9. выполнять каприз
10. выполнять команду
11. выполнять нагрузку
12. выполнять наказ
13. выполнять норму
14. выполнять обещание
15. выполнять обязанности
16. выполнять обязательства
17. выполнять план
18. выполнять поручение
19. выполнять правила
20. выполнять приказ
21. выполнять приказание
22. выполнять программу
23. выполнять просьбу
24. выполнять работу
25. выполнять распоряжение
26. выполнять решение
27. выполнять роль
28. выполнять совет
29. выполнять требование
30. выполнять указание
31. выполнять упражнение
32. выполнять условие
33. выполнять установку
34. выполнять функцию

Задача первого эксперимента заключалась в том, чтобы определить, влияет ли учёт частеречной принадлежности коллоката ключевого слова (глагола) на качество выделения устойчивых словосочетаний. Исследовались диапазоны [-1, 1], [-2, 2], [-3, 3]. Второй эксперимент состоял в изучении того, как ширина контекста влияет на качество выделения коллокаций с учётом частеречной принадлежности коллоката. Исследовались диапазоны от 1 до 5 слов справа от глагола.

В ходе экспериментов было обнаружено, что система IntelliText выделяет словосочетания, которые не зафиксированы в вышеперечисленных словарях, но, тем не менее, могут быть отнесены к разряду устойчивых словосочетаний. Для решения этой проблемы была проведена экспертная оценка: экспертам предлагалось определить, какие из выделенных программой IntelliText словосочетаний являются коллокациями. На основе полученных данных список устойчивых глагольно-именных словосочетаний был расширен. Например, в расширенный список глагольно-именных словосочетаний для глагола выполнять вошли следующие словосочетания: *выполнять волю, выполнять движение, выполнять действие, выполнять контракт, выполнять маневр, выполнять миссию, выполнять обработку, выполнять операцию, выполнять пожелание, выполнять предписание, выполнять прыжок, выполнять расчёт, выполнять рекомендацию, выполнять соглашение, выполнять трюк*.

Для оценки качества выделения устойчивых словосочетаний для каждого случая были подсчитаны метрики точности (Precision), полноты (Recall) и гармонического среднего (F-measure), вычисляемых по следующим формулам:

$Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}$, где

D_{rel} – множество релевантных словосочетаний из списка устойчивых словосочетаний;

D_{retr} – множество всех словосочетаний, найденных системой.

$Recall = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}$, где

D_{rel} – множество релевантных словосочетаний из списка устойчивых словосочетаний;

D_{retr} – множество всех словосочетаний, найденных системой.

$F = (\beta^2+1)\frac{Precision \times Recall}{\beta^2 Precision + Recall}$, где $\beta = 2$ (приоритет отдаётся полноте).

3. Эксперименты

В таблице 1 представлены результаты первого эксперимента (исследование влияния учёта частеречной принадлежности коллокатов на выделение коллокаций) для глагола *выполнять*. В первом столбце таблицы указывается ранг кандидата, в остальных столбцах приведены кандидаты, выделенные системой IntelliText для разных контекстов, с учётом и без учёта частеречной принадлежности предполагаемого коллоката. Белым фоном отмечены слова, которые, согласно эталонному списку выделенных нами словосочетаний, являются коллокатами для глагола выполнять

Таблица 1. Выделение коллокатов для глагола *выполнять* с помощью системы IntelliText (с учётом и без учёта частеречной принадлежности коллоката)

ранг	[-1, 1]		[-2, 2]		[-3, 3]	
	без учёта части речи	**сущ. в вин.п.**	**без учёта части речи**	**сущ. в вин.п.**	**без учёта части речи**	**сущ. в вин.п.**
1.	организационные	приказание	организационные	приказание	перевыполнять	приказание
2.	пунктуально	функция	перевыполнять	функция	беспрекословно	функция
3.	беспрекословно	наказ	беспрекословно	поручение	неукоснительно	поручение
4.	неукоснительно	поручение	неукоснительно	наказ	функция	наказ
5.	добросовестно	задание	добросовестно	задание	добросовестно	задание
6.	приказание	приказ	приказание	обязанность	приказание	обязанность
7.	функция	упражнение	функция	предписание	поручение	предписание
8.	скрупулёзно	указание	поручение	приказ	задание	обещание
9.	безукоризненно	заказ	наказ	обещание	обязанность	заказ
10.	наказ	предписание	задание	заказ	упражнение	приказ
11.	безропотно	обязанность	упражнение	упражнение	обещание	упражнение
12.	поручение	роль	обещание	обязательство	заказ	обязательство
13.	приказ	директива	обязанность	указание	предписание	миссия
14.	задание	обещание	приказ	директива	приказ	указание
15.	исправно	рейс	заказ	миссия	обязательство	директива
16.	упражнение	инструкция	предписание	трюк	указание	предназначение
17.	возложить	завет	обязательство	роль	миссия	долг
18.	послушно	обязательство	указание	долг	возложить	трюк
19.	интернациональный	распоряжение	возложить	распоряжение	долг	роль
20.	заказ	задача	послушно	предназначение	защитный	распоряжение
21.	указание	маневр	миссия	завет	роль	задача
22.	предписание	трюк	трюк	задача	задача	завет
23.	управленческий	требование	защитный	пожелание	распоряжение	пожелание
24.	обязанность	приветствие	долг	приветствие	рейс	каприз
25.	обещание	миссия	роль	рейс	инструкция	рейс
26.	качественно	пожелание	распоряжение	инструкция	воинский	инструкция
27.	роль	полёт	задача	каприз	обязанный	приветствие
28.	защитный	заповедь	рейс	маневр	требование	просьба
29.	призванный	норма	эффективно	требование	служебный	требование
30.	эффективно	рекомендация	инструкция	рекомендация	успешно	заповедь
31.	обязательство	обряд	воинский	просьба	просьба	рекомендация
32.	успешно	план	обязанный	заповедь	работа	бросок
33.	рейс	воля	требование	обряд	строго	маневр
34.	старательно	просьба	успешно	воля	норма	обряд
35.	инструкция	долг	служебный	полёт	операция	норма
36.	обязанный	блок	рекомендация	норма	полёт	воля
37.	распоряжение	работа	строго	ритуал	боевой	работа
38.	строго	стрельба	просьба	работа	способный	ритуал
39.	задача	посадка	полёт	план	честно	полёт
40.	ежедневно	команда	работа	стрельба	план	план
41.	чётко	операция	способный	перевозка	воля	посадка
42.	полёт	удар	чётко	посадка	команда	перевозка
43.	честно	готовность	норма	операция	объём	монтаж
44.	способный	условие	боевой	команда	позволять	классификация
45.	требование	танец	честно	прыжок	одновременно	стрельба

46.	самостоятельно	поворот	воля	заявка	удар	операция
47.	боевой	множество	операция	блок	определённый	назначение
48.	служебный	программа	план	спектр	сложный	команда
49.	важнейший	фигура	позволять	назначение	приходиться	процедура
50.	отказываться	желание	одновременно	программа	правило	прыжок
51.	броситься	решение	команда	процедура	действие	заявка
52.	норма	ряд	объём	ремонт	важный	спектр
53.	одновременно	правило	удар	установка	различный	блок
54.	фактически	способность	определённый	желание	движение	желание
55.	позволять	расчёт	приходиться	удар	предприятие	нагрузка
56.	план	закон	сложный	условие	свой	программа
57.	приходиться	контроль	прямой	фигура	любой	установка
58.	работа	возможность	важный	цикл	программа	действие
59.	определённый	движение	различный	обслуживание	социальный	удар
60.	воля	шаг	свой	правило	организация	ремонт
61.	просьба	действие	правило	действие	должен	обслуживание
62.	команда	срок	любой	расчёт	условие	правило
63.	долг	качество	движение	контракт	решение	комплекс
64.	способность	сила	программа	готовность	который	условие
65.	объём	-	должен	решение	лишь	фигура
66.	свой	-	социальный	объём	также	решение
67.	операция	-	действие	танец	всегда	цикл
68.	различный	-	предприятие	поворот	они	расчёт
69.	сложный	-	условие	множество	весь	объём
70.	удар	-	организация	контроль	надо	контракт
71.	любой	-	решение	комплекс	каждый	готовность
72.	специальный	-	основной	приём	по	движение
73.	должен	-	государственный	ряд	этот	контроль
74.	важный	-	закон	услуга	или	танец
75.	поставить	-	также	движение	другой	приём
76.	продолжать	-	всегда	закон	мочь	поворот
77.	условие	-	лишь	соглашение	только	услуга
78.	прийтись	-	который	способность	его	множество
79.	точно	-	она	исследование	тот	способность
80.	следующий	-	надо	разработка	,	закон
81.	лишь	-	весь	шаг	и	ряд
82.	также	-	можно	проект	не	течение
83.	всегда	-	мочь	товар	они	срок
84.	решение	-	только	анализ	же	урок
85.	начать	-	этот	техника	)	исследование
86.	весь	-	другой	возможность	она	соглашение
87.	они	-	его	очередь	мы	шаг
88.	надо	-	или	течение	то	разработка
89.	можно	-	по	срок	.	реализация
90.	мочь	-	тот	совет	но	проект
91.	который	-	не	число	:	техника
92.	его	-	он	организация	"	очередь
93.	только	-	она	часть	что	счёт
94.	тот	-	мы	счёт	быть	круг
95.	этот	-	и	время	в	возможность
96.	не	-	,	качество	а	совет

97.	она	-	)	право	как	число
98.	быть	-	быть	положение	"	время
99.	он	-	"	процесс	с	часть
100.	,	-	.	день	на	день
101.	и	-	в	дело	–	дело

Таблица 2 позволяет установить, что учёт частеречной принадлежности предполагаемого коллоката повышает точность, полноту и F-меру при выделении коллокаций. Таким образом, результаты первого эксперимента позволяют сделать вывод о том, что учёт частеречной принадлежности коллоката повышает качество выделения глагольно-именных словосочетаний.

Таблица 2. Оценка результатов выделения глагольно-именных словосочетаний для глагола *выполнять*

Диапазон	**Точность**	**Полнота**	**F-мера**
[-1, 1] (без учёта ч.р.)	0,2871	0,5918	0,4882
[-1, 1] (с учётом ч.р.)	0,625	0,8163	0,7692
[-2, 2] (без учёта ч.р.)	0,3366	0,6939	0,5724
[-2, 2] (с учётом ч.р.)	0,4653	0,9592	0,7912
[-3, 3] (без учёта ч.р.)	0,3168	0,6531	0,5387
[-3, 3] (с учётом ч.р.)	0,4851	1	0,8249

Анализ словосочетаний, выделенных системой IntelliText в первом эксперименте, показал, что левый контекст не содержит релевантных коллокатов для исследуемых нами глаголов. Это можно объяснить тем, что в русском языке прямое дополнение (существительное в винительном падеже) в большинстве случаев находится в постпозиции по отношению к глаголу. Поэтому во втором эксперименте мы ограничились правым контекстом: исследовались диапазоны от 1 до 5 слов справа от глагола.

Результаты второго эксперимента (исследование влияния ширины контекста на выделение коллокаций на примере глагола *выполнять*) представлены в таблице 3. В первом столбце указывается ранг кандидата, в остальных столбцах указаны выделенные системой IntelliText существительные в винительном падеже для контекстов от 1 до 5 слов справа от глагола. Аналогично предыдущей таблице, белым цветом выделены те существительные, которые являются коллокатами для глагола выполнять.

Согласно результатам второго эксперимента, с увеличением ширины контекста показатели точности и F-меры понижаются. Показатель полноты повышается и достигает максимальной величины при диапазонах [0, 4] и [0, 5] (см. График 1).

Анализ данных эксперимента позволяет сделать следующие наблюдения. С одной стороны, увеличение ширины контекста помогает выделить разрывные коллокации (словосочетания, компоненты которых разделены другим словом).

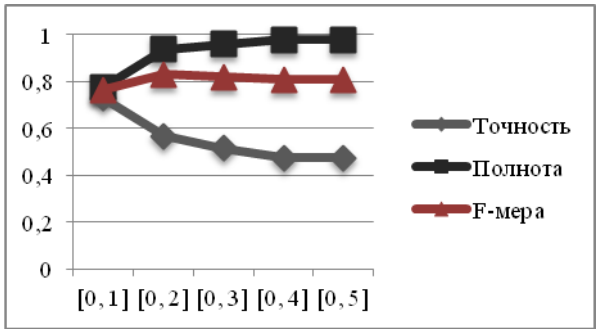

График 1. Оценка результатов выделения словосочетаний для глагола *выполнять*

Ниже приведены типовые примеры разрывных коллокаций глагола *выполнять* из корпуса текстов RNC2010-MOCKY, выделяемые для разных контекстов.

Контекст = [0, 2].

…он продолжает работать, **выполняя** *основные* **функции**, которые не требуют большого потребления…

…выяснение, насколько Америка способна **выполнять** *эту* **роль** гипердержавы. И там нет уверенности…

Контекст = [0, 3].

…лётчики и подводники, **выполняющие** ***особые*** *правительственные* **задания**, офицеры военной разведки…

…Его ум позволяет радоваться, **выполняя** *приказы и* **указания**. Наверно, это круто…

Контекст = [0, 4].

…в приоритетном порядке **выполнять** *взятые по конвенции* **обязательства.** Президент США…

…постоянно **выполняло** *и перевыполняло государственный* **план**, занимая первое место…

Контекст = [0, 5].

…Ваша задача научиться **выполнять** *любые атакующие и защитные* **движения** так, будто они уже достигли…

...посетителей на стенде «Линии График», **выполнявших** *какие-то только им понятные* **действия**, привлекало…

С другой стороны, увеличение ширины контекста и отсутствие учёта синтаксических связей между словами в тексте приводит к неправильному выделению коллокатов.

В качестве примеров приведены основные ошибки при выделении коллокатов глагола *выполнять* из корпуса текстов RNC2010-MOCKY.

Контекст = [0, 3].

…Старайтесь с первых же шагов в обучении **выполнять** *предоставленные методы* **действия** без предварительной подготовки…

В данном случае неверно выделено словосочетание **выполнять действия***, так как существительное* **действия** *является зависимым по отношению к существительному* **методы***, а не к глаголу* **выполнять.**

Контекст = [0, 4].

…Свою работу эксперты наши **выполняют** *добросовестно, а дальнейшие* **действия** - кто в дальнейшем будет манипулировать…

Системой IntelliText в качестве коллоката для глагола **выполняют** *выделено существительные* **действия***, которое является частью второго простого предложения в составе сложного.*

…Можно заставить подчинённых **выполнять** *работу, отдав соответствующее* **распоряжение**, но такой процесс…

Неправильно выделено словосочетание **выполнять распоряжение***, так как существительное* **распоряжение** *является частью деепричастного оборота и является зависимым по отношению к деепричастию* **отдав.**

Для решения проблемы выделения ошибочных словосочетаний предполагается разработка собственной программы выделения коллокаций, сочетающей статистические (меры ассоциации) и синтаксические средства (учёт синтаксических связей между компонентами словосочетаний).

Аналогичные эксперименты были проведены для глаголов *нарушать* и *принимать*. Анализ полученных результатов позволяет сделать вывод о том, что, как и в случае с глаголом *выполнять*, точность, полнота и F-мера при выделении коллокаций для глаголов *нарушать* и *принимать*, а, следовательно, и качество выделения глагольно-именных словосочетаний выше при учёте части речи предполагаемого коллоката.

3. Выводы

На основе проведённых экспериментов можно сделать следующие выводы:

1. Качество выделения устойчивых глагольно-именных словосочетаний с учётом части речи коллоката выше, чем без учёта его частеречной принадлежности;
2. Увеличение ширины контекста при учёте частеречной принадлежности коллоката, в основном, повышает полноту, но понижает точность и F-меру выделяемых словосочетаний;
3. Увеличение ширины контекста при учёте части речи коллоката имеет неоднозначный характер: с одной стороны, это средство позволяет выделять разрывные словосочетания; с другой стороны, оно приводит к возникновению ошибок при выделении коллокаций, что говорит о необходимости учёта не только частеречной принадлежности коллоката, но и синтаксических связей между компонентами словосочетаний.
4. Результаты данных экспериментов обнаруживают, что статистическая мера MI является эффективной мерой ассоциации и ранжирования при выделении глагольно-именных словосочетаний. Ранги предполагаемых коллокатов свидетельствуют об их отношении к коллокациям в реальной речи: высокий ранг коллоката соответствует его широкому употреблению в качестве компонента устойчивых словосочетаний; низкий ранг коллоката, наоборот, говорит о том, что данный коллокат чаще всего входит в состав свободных словосочетаний.

Таким образом, анализ и сравнительная оценка методов автоматического выделения словосочетаний позволяет сделать вывод о необходимости комплексного подхода к решению данной задачи, иными словами, совместного использования статистического и синтаксического методов поиска и выделения коллокаций.

Литература

[1] Дерибас В.М. Устойчивые глагольно-именные словосочетания русского языка. М., 1983.
[2] Словарь русского языка: В 4 т. (МАС) / Под ред. А.П. Евгеньевой. – 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.
[3] Словарь сочетаемости слов русского языка./Под ред. П.Н. Денисова, В.В. Морковкина. – 2-е изд., испр. М.: Русский язык, 1983.
[4] Church, K., Gale, W., Hanks, P. and Hindle, D. 1991. Using Statistics in Lexical Analysis. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. New Jersey, Lawrence Erlbaum. P. 115-164.
[5] Evert, S. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, University of Stuttgart, 2004.
[6] Pečina P. Lexical Association Measures: Collocation Extraction. Praha, 2009.

Comparing Methods of Automatic Verb-Noun Collocation Extraction

Svetlana Koshcheeva

The paper describes a set of experiments the aim of which is to compare several approaches to automatic verb-noun collocation extraction. The main subjects under observation are the impact of span size and POS-filtering on the quality of collocation extraction. The experiments have shown that collocations lists extracted by means of POS-filtering are significantly more precise than those obtained without POS-filtering, whereas the extension of a span size has an ambiguous effect. On the one hand, it enables the extraction of distant collocates, but on the other hand it results in erroneous collocates, which leads therefore to consider the use of syntax-based approach for verb-noun collocation extraction.