Дистрибутивная модель сочетаемостных ограничений глаголов[1]

В. И. Пекар
Башкирский государственный педагогический университет

Сочетаемостные ограничения глагола как лингвистическое понятие обозначает семантические ограничения, которые должны удовлетворяться словами, синтаксически зависимыми от глагола. Компьютерное моделирование сочетаемостных ограничений может иметь важное значение для различных задач ОЕЯ: машинного перевода, дизамбигуации многозначных слов, интерпретации анафоры, семантического аннотирования текста. В настоящей работе проводится анализ методов автоматического отбора дистрибутивных признаков, наиболее релевантных для моделирования сочетаемостных ограничений глагола.

1. Компьтерное моделирование сочетаемостных ограничений

Сочетаемостные (или *селекционные*) *ограничения* глагола как лингвистическое понятие обозначает семантические ограничения, которые должны удовлетворяться существительными, синтаксически зависимыми от глагола. Например, прямые дополнения при глаголе *есть* должны иметь семантический компонент «еда», а при глаголе *пить* – компонент «жидкость»[2]. Компьютерное моделирование сочетаемостных ограничений может иметь важное значение для различных задач обработки естественного языка, в которых необходимо предсказать появление определенного слова или определенной категории слов в потоке речи при условии присутствия в ней других слов. Так, модель сочетаемостных ограничений может использоваться при машинном переводе для контекстно-зависимого выбора переводных эквивалентов: наиболее подходящий вариант перевода многозначного слова может быть выбран, если известен перевод другого слова, от которого многозначное зависит синтаксически. Для семантического аннотирования текста, необходимого, например, для обучения системы извлечения информации, сочетаемостные ограничения слов могут использоваться для предсказания наиболее вероятного семантического класса зависимых от них слов.

---

[1] Работа выполнена в рамках исследовательского проекта, финансируемого грантом РФФИ № 03-06-80008.

[2] В настоящее время существует тенденция говорить и о сочетаемостных ограничениях, которые слова, принадлежащие к другим частям речи, накладывают на свое лексическое окружение, например, об ограничениях, накладываемых существительными на глаголы и прилагательные (см., например, [1,3]). В данной работе речь идет только о сочетаемостных ограничениях глаголов.

Основная проблема компьютерного моделирования сочетаемостных ограничений слова может быть сформулирована как разработка такого способа их абстрактного представления, на основании которого можно было бы оценить степень сочетаемости этого слова с любым другим. Многие из методов, разработанных в компьютерной лингвистике, представляют сочетаемостные ограничения как семантический класс слов. Принадлежность слов этому классу трактуется как их способность удовлетворить моделируемым сочетаемостным ограничениям. Например, ограничения, накладываемые глаголом *есть* на прямые дополнения, можно представить в виде класса слов, объединенных понятием «еда». Существуют две точки зрения на то, какие классы слов должны использоваться для этой цели. Наиболее популярным подходом является использование иерархически организованных классов слов в тезаурусе, например, синонимические группы WordNet. Суть этого подхода заключается в нахождении для аргументной позиции глагола класса существительных с наиболее оптимальной степенью абстрактности. Такой класс выбирается с помощью данных корпуса: в корпусе находятся существительные, которые употреблялись в этой аргументной позиции. Затем определяется тот класс, гипонимы которого включают в себя эти существительные. Существует целый ряд методик нахождения такого класса: selectional association measure [8], tree-cut model [5], скрытые модели Маркова [1]. Основные достоинства этого подхода обычно видятся в его способности приобрести репрезентации сочетаемостных ограничений их разряженных и шумных данных. Его очевидный недостаток, однако, заключается в том, что существительные – как те, на основании которых строится модель, так и те, чья роль в качестве аргумента впоследствии проверяется – должны присутствовать в тезаурусе. Это требование трудно удовлетворить для слов-терминов, принадлежащих специальным предметным областям, а также для имен собственных, т.е. для лексических единиц, составляющих значительную часть вокабуляра корпуса. Другой недостаток этого подхода заключается в лежащем в его основе предположении о том, что фиксированное число статичных, дизъюнктивных синонимичных групп тезауруса способно адекватно отразить сочетаемостные ограничения *любого* глагола.

Во втором подходе, альтернативном тезаурусному, сочетаемостные ограничения представляются в виде динамических семантических классов, построенных исключительно на основе данных корпуса (см., например, [2-4]. Здесь классы, служащие представлением сочетаемостных ограничений, тоже состоят из существительных, которые употреблялись в корпусе в качестве аргументов глагола. Однако, их объединение в классы и определение классовой принадлежности новых слов осуществляется на основе дистрибутивной схожести между словами, т.е. схожести в контекстах их употребления. Таким образом оказывается возможным, во-первых, использовать контекстно-релевантные классы слов, принадлежность которым может быть градуальной, и, во-вторых, избавиться от требования присутствия всех существительных в тезаурусе.

Данная работа посвящена моделированию сочетаемостных ограничений глаголов с помощью второго метода. Цель работы заключается в изучении возможностей повысить качество модели путем отбора дистрибутивных признаков слов. В работе изучается влияние различных методов отбора признаков на способность компьютерной системы предсказать степень сочетаемости слов, а также рассматривается их взаимодействие со способом измерения дистрибутивной схожести между словами.

2. Дистрибутивная модель

Дистрибутивный подход строит репрезентацию сочетаемостных ограничений глагола v путем абстрагирования от дистрибутивных данных множества существительных N_v, употреблявшихся при v в качестве аргумента. Сначала каждое $n \in N_v$ представляется в виде

признакового вектора $<f_1, f_2, \ldots, f_z>$, где признаки соответствуют контекстам употребления n. Значение каждого признака – условная вероятность $p(f_i|n)$, вычисленная из данных корпуса. Семантическая совместимость глагола v и любого данного существительного m численно выражается как условная вероятность $p(m|v)$, которая вычисляется следующим образом. Среди существительных N_v сначала отбирается k наиболее схожих с m путем сравнения их дистрибутивных репрезентаций. Эти существительные составляют группу ближайших соседей m N'_v. Затем $p(m|v)$ подсчитывается как среднее известных из корпуса $p(n \in N'_v|v)$, каждая из которых взвешена показателем схожести между n и m.
Для того, чтобы некое существительное удовлетворило сочетаемостным ограничениям глагола необходимо наличие в его значении лишь некоторых семантических компонентов, остальная часть его семантики нерелевантна для оценки его совместимости с глаголом. Поэтому представляется, что и при оценке дистрибутивной схожести между существительными N_v и m лишь только определенная часть их признаков релевантна для вычисления $p(m|v)$. В настоящей работе проверяется гипотеза о том, качество модели сочетаемостных ограничений может быть повышено, во-первых, путем отбора наиболее информативных признаков слов n на основании их распределения по репрезентациям всех существительных корпуса, и, во-вторых, использованием такой меры дистрибутивной схожести между этими существительными и m, которая подсчитывает число отобранных признаков у m, игнорируя все остальные его признаки.

3. Методы селекции признаков

В машинном обучении автоматический отбор наиболее полезных признаков объектов производится на основе распределения признаков по репрезентациям отдельных объектов или по их классам. Наиболее информативными считаются те признаки, которые наиболее неравномерно распределены. Процедура отбора заключается в вычислении некой меры релевантости для каждого признака, являющейся функцией от его частоты появления у различных объектов или классов объектов. Затем осуществляется отбор признаков – либо из всего пространства признаков, либо из признаков отдельных объектов удаляется некоторая доля наименее информативных. Для решения текущей задачи (например, построения классификатора или кластерного анализа) используются только оставшиеся признаки.
Для отбора признаков существительных в настоящей работе использовались три метода оценки ассоциации между двумя словами, часто используемые в статистической ОЕЯ: хи-квадрат, точный тест Фишера и Gain Ratio. Все три метода вычисляют показатель релевантности признака f для слова n из данных таблицы сопряженности с двумя строками и двумя столбцами (см.таб.1), где числа в полях соответствуют частотам появления f и $\bar{f}$ в репрезентациях n и $\bar{n}$ ($\bar{f}$ и $\bar{n}$ означают “не-f” и “не-n”, соответственно).

	f	$\bar{f}$
n	10	1
$\bar{n}$	1	100

Таб.1. Данные, используемые для вычисления меры релевантости признака f для слова n.

3.1. Хи-квадрат

Хи-квадрат является статистическим тестом, широко используемым в различных экспериментальных науках для оценки достоверности отличия наблюдаемых результатов от ожидаемых. В качестве меры релевантности признака f для существительного n можно использовать показатель χ^2 зависимости между наблюдаемой частотой $fr(f,n)$ появления f у n

и ожидаемой частотой *fr'*(*f,n*). Последняя вычисляется на основании предположения, что *f* случайно появляется у *n*, подсчитываемая как произведение *fr*(*f*) и *fr*(*n*), разделенное на число всех пар "признак-слово", имеющихся в корпусе:

$$fr'(f,n) = \frac{fr(f) \cdot fr(n)}{\sum_{g \in F} \sum_{h \in N} fr(g,h)} \quad (1)$$

Показатель хи-квадрат (χ^2) вычисляется как

$$\chi^2(f,c) = \sum_{g \in \{f,\bar{f}\}} \sum_{h \in \{n,\bar{n}\}} \frac{(fr(g,h) - fr'(g,h))^2}{fr'(g,h)} \quad (2)$$

Наибольшие показатели χ^2 говорят о наибольшей зависимости между *f* и *n*, и следовательно о релевантности *f* для представления семантики *n*.

3.2. Точный тест Фишера

Как и хи-квадрат, точный тест Фишера используется для оценки зависимости между двумя событиями. В отличие от теста хи-квадрат точный тест Фишера не зависит от предположения о том, что данные характеризуются распределением хи-квадрат. Поэтому тест Фишера часто применяется как альтернатива ему, в частности в случаях, когда число наблюдений в отдельном поле таблицы сопряженности меньше 6. Тест Фишера подсчитывается следующим образом. Сначала суммируются частоты по строкам и столбцам в таблице, затем вычисляются условные вероятности получения всех возможных распределений частот в таблице при условии, что их сумма по строкам и столбцам соответствует наблюдаемой. Сумма условных вероятностей, которые меньше или равны условной вероятности получить наблюдаемое распределение, составляет значение *P*-оценки, вероятности того, что два события независимы.

Мера релевантности признака, вычисляемая с помощью теста Фишера (ТФ), подсчитывается следующим образом (процедура согласно [7]). Сначала вычисляется вероятность получения такого распределения *f* и *n*, при котором они никогда не появлялись одновременно, при условии, что сумма частот в таблице соответствует наблюдаемым. Затем вычисляется значение *P*-оценки того, что *f* и *n* независимы при таком распределении. Поскольку фактически *f* и *n* проявляются одновременно, то высокая вероятность их независимости, в случае если бы они никогда не проявлялись одновременно, может служить мерой их ассоциации. Другими словами, чем больше степень ожидаемой независимости между событиями при их фактическом одновременном проявлении, тем выше ассоциация между ними.

3.3. Gain Ratio

Gain Ratio является вариантом меры *Information Gain*, широко используемым в машинном обучении для оценки релевантности признаков объектов. *Information Gain* (IG) как понятие из теории информации обозначает количество бит, который одна переменная содержит о другой. IG между признаком *f* и словом *n* выражает разницу между энтропией *n* при наличии *f* и энтропией *n* при отсутствии *f*. IG вычисляется следующим образом:

$$IG(f,c) = \sum_{g \in \{f,\bar{f}\}} \sum_{h \in \{n,\bar{n}\}} p(g,h) \log \frac{p(g,h)}{p(g)p(h)} \quad (3)$$

Недостаток IG заключается в том, что значение IG повышается не только по мере увеличения ассоциации между *f* и *n*, но и по мере увеличения энтропии *f*. Gain Ratio (GR) избавляется от этого недостатка с помощью деления IG на энтропию *f*:

$$GR(f,n) = \frac{IG(f,n)}{-\sum_{g\in\{f,\bar{f}\}} p(g)\log p(g)} \qquad (4)$$

4. Данные экспериментов

В экспериментах проверялась способность компьютерной программы правильно предсказать степень сочетаемости глагола с существительными, выступающими при нем в качестве дополнений. Для этой цели из данных Британского Национального Корпуса были автоматически собраны пары “глагол-дополнение” и “прилагательное-существительное”. После удаления существительных, употреблявшихся в менее чем 5 разных контекстах, эти данные использовались для построения дистрибутивных репрезентаций существительных. Они составили “тренировочную часть” данных. Затем была создана “тестовая часть” данных. Из пар “глагол-дополнение” были отобраны глаголы, чьи селекционные ограничения моделировались в ходе экспериментов. Чтобы исключить слишком частые пары, которые могут быть устойчивыми словосочетаниями, и слишком редкие, необычные употребления слов, для этой цели были отобраны те пары, которые появлялись в корпусе более 5 и менее 300 раз. Были также исключены пары, в которых глагол и/или существительное входит в число 100 наиболее частотных слов, чтобы исключить из экспериментов слова со слишком абстрактной семантикой. Из оставшихся пар методом случайной выборки было отобрано 1000 пар. Для каждой пары “глагол-дополнение” среди существительных, для которых были построены дистрибутивные репрезентации, случайно выбиралось одно, имеющее примерно то же число признаков, что употребляемое при глаголе. В таблице 2 приводятся примеры полученных таким образом троек слов, из которых одно существительное является допустимым аргументом глагола, а второе, случайно выбранное, скорее всего недопустимо в качестве его аргумента. 1000 отобранных для тестов пар глагол-“правильное” дополнение были удалены из “тренировочных данных”.

Глагол	*сущест-ное1*	*сущест-ное2*
acquire	momentum	ease
aid	reader	evening
apologise_for	remark	lecture
approach	door	eye
arrive_at	evaluation	diagram
assume	significance	tape
award	championship	validity
befit	rank	delight
blame_for	fire	lip
breathe	air	degree

Таб 2. Примеры троек слов, использовавшихся в качестве тестовых данных.

Процедура экспериментов заключалась в следующем. Из каждой тройки слов из тестовых данных, строилась модель сочетаемостных ограничений глагола на основе данных тренировочной части. Затем оценивалась вероятность сочетаемости глагола с каждым из двух существительных и подсчитывалась *аккуратность* решений о сочетаемости слов:

$$A = \frac{1}{N}(c + \frac{u}{2}) \qquad (5)$$

где N – число тестовых троек слов, c – число случаев, когда оцененная вероятность сочетания с глаголом для «правильного» существительного больше вероятности для «неправильного», а u – число случаев, когда оцененные вероятности для обоих существительных одинаковы.

5. Результаты

Излагаемые ниже результаты были получены путем деления 1000 тестовых троек слов на 10 равных частей, оценки аккуратности внутри каждой из них и подсчета ее среднего показателя по 10 частям. В ходе экспериментов по отбору наиболее релевантных признаков варьировался масштаб сокращения числа признаков у существительного (отбирались 10, 20, …, 90% его наиболее информативных признаков), а также параметр k в используемом методе моделирования селекционных ограничений, т.е. число ближайших соседей существительного m (см. раздел 2). При измерении схожести между дистрибутивными репрезентациями слов использовалась функция Jensen-Shannon Divergence, которая по результатам исследования [2] является наиболее предпочтительной. Изучаемые методы отбора признаков сравнивались между собой, а также с аккуратностью, достигаемой без предварительного отбора признаков, а также при их случайном отборе.

На рис.1 слева показан график зависимости аккуратности от масштаба селекции признаков, выполняемых с помощью трех анализируемых методов, а также путем случайного отбора (для каждого из методов приводятся результаты для наиболее оптимального показателя k: для χ^2 – 50, для ТФ – 20, для GR – 100, при случайном отборе – 100). Без отбора признаков (т.е., при использовании 100% признаков) аккуратность составила 0,708 при k=3.

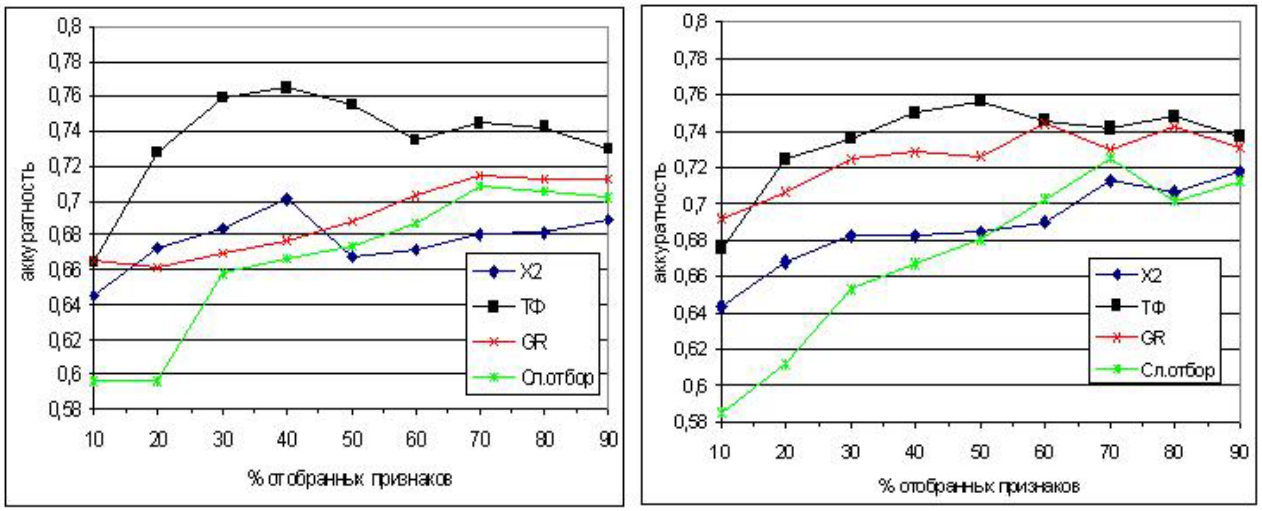

Рис.1. Зависимость показателя аккуратности от масштаба селекции признаков, производимой с помощью анализируемых методов. График слева показывает результаты при использовании метода оценки схожести Jensen-Shannon Divergence, график справа – при использовании метода Overlap Coefficient.

Как видно из этих результатов, отбор признаков методом ТФ позволяет достичь более высокой аккуратности по сравнению с χ^2, GR и случайным отбором признаков. Отличия в наилучших результатах для ТФ (при k=20, использование 40% признаков слов) от наилучших результатах при случайном отборе признаков (k=100, 70% признаков) и от результатов, полученных без отбора признаков, статистически достоверны при их оценке с помощью

парного t-теста на уровне α=0,001. Однако различия между результатами, достигнутыми методами χ^2 и GR, и результатами, полученных путем случайного отбора признаков и без их отбора, статистически недостоверны.
Далее проверялась гипотеза о том, что после отбора признаков наиболее эффективным способом измерения схожести между словами является такой, который проверяет присутствие у тестового слова признаков, отобранных на основе тренировочных данных, и игнорирует наличие у него всех остальных. Для этой цели использовалась функция Overlap Coefficient [6], подсчитывающая схожесть между двумя признаковыми репрезентациями *n* и *m* как степень включения меньшей из них в большую:

$$OC(n,m) = \frac{|F_n \cap F_m|}{\max(|F_n|,|F_m|)} \quad (6)$$

Вместо числа признаков меньшей репрезентации в знаменателе мы использовали число отобранных признаков тренировочного слова.
Результаты экспериментов с использованием этой меры дистрибутивной схожести (рис.1., график справа) говорят о том, что этот способ измерения схожести между словами действительно несколько повышает аккуратность для χ^2 (на 1,7%) и для GR (на 3%, достоверно при α=0,05), но не для ТФ. Эффективность моделей также повысилась и при случайном отборе признаков (на 1,7%). Аккуратность χ^2 не превысила аккуратности, достигаемой без отбора признаков (последняя составила 0,732 при *k*=7), и при их случайном отборе. Аккуратность GR была выше аккуратности без отбора признаков лишь на 1,1%, и выше аккуратности при случайном отборе на 2% (в обоих случаях различия недостоверны). ТФ продемонстрировал достоверно большую эффективность по сравнению со случайным отбором признаков (α=0,05), но не по сравнению с использованием 100% признаков.

6. Заключение

В данной работе были рассмотрены возможности повышения качества дистрибутивных моделей сочетаемостных ограничений глагола путем отбора наиболее релевантных признаков для их представления. Предпочтительным способом отбора оказался метод, основанный на измерении зависимости между словом и его признаком с помощью точного теста Фишера. Этот метод показал большую эффективность, чем методы хи-квадрат и Gain Ratio. Повышая качество моделей, тест Фишера позволяет вдвое уменьшить размерность репрезентаций сочетаемостных ограничений. В данной работе был также предложен новый способ измерения схожести между репрезентацией сочетаемостных ограничений и существительным, чью семантическую совместимость с глаголом необходимо проверить. Хотя при использовании разных методов отбора предложенный способ показал лучшие результаты, чем способ, обычно используемый при дистрибутивном моделировании сочетаемостных ограничений, разница между ними была достоверна лишь для одного метода отбора признаков.

Литература

1. Abney S. and Light M. Hiding a semantic class hierarchy in a Markov model. // Proceedings of the ACL workshop on Unsupervised Learning in NLP, 1999. pp.1-8.
2. Dagan I., Pereira F., Lee L. Similarity-based models of co-occurrence probabilities. Machine Learning, 34 (1-3), 1999. pp.43-69.

3. Lapata M., McDonald S., Keller F. Determinants of noun-adjective plausibility. // Proceedings of EACL, 1999. pp.30-36.
4. Lee L. On the effectiveness of the skew divergence for the statistical language analysis. Artificial Intelligence and Statistics, 1999. pp.65-72.
5. Li H. and Abe N. Generalizing case frames using a thesaurus and the MDL principle. Computational Linguistics. 24 (2), 2001. pp.217-244.
6. Manning C. and Schuetze H. Foundations of statistical natural language processing. Cambridge, MA: The MIT Press, 1999.
7. Pedersen T. Fishing for exactness. // Proceedings of SCSUG, 1996. pp.188-200.
8. Resnik P. Selectional preferences and sense disambiguation. // Proceedings of the ACL SIGLEX workshop on Tagging text with lexical semantics: why, what, and how?, 1997. pp.52-57.