

Две стратегии взвешивания признаков при дистрибутивной классификации слов

В.И.Пекар

Башкирский государственный педагогический университет

1. Введение

Семантические ресурсы (электронные тезаурусы, лексиконы и др.) сейчас являются важной частью многих приложений ОЕЯ, таких как вычленение информации из текста, вопросно-ответные системы или поиск информации по семантическим критериям. Быстрая разработка таких ресурсов и адаптация существующих к новым областям знания может быть возможна при использовании методов машинного обучения. В настоящее время появляется значительное число работ, направленных на поиск возможностей автоматически сформировать некие репрезентации значений слов на основе данных корпуса и затем использовать их для распределения новых слов по синонимическим группам тезауруса с помощью методов автоматической классификации или кластеризации (см., например, Hearst, 1992; Roark&Charniak, 1998; Lin 1998).

Настоящее исследование посвящено семантической классификации слов на основе их дистрибутивных репрезентаций. Дистрибутивный подход в представлении значения слова, благодаря своей способности полностью автоматизировать обработку лексического материала, сейчас широко используется для различных задач, связанных с семантическими ресурсами. Так, есть опыт его применения для построения новых семантических ресурсов (Lin 1998; Caraballo, 1998), пополнения существующих ресурсов новыми лексическими единицами (Tokunaga et al., 1997; Alfonseca&Manandhar, 2002; Pekar&Staab, 2002), новыми значениями для многозначных слов, уже имеющих в базе данных (Lin&Pantel, 2002).

Согласно этому подходу, значение слова представляется в виде вектора признаков, где каждый вектор соответствует отдельному контексту употребления слова, обнаруженному в корпусе (как правило, в качестве контекста выступает синтаксически связанное с ним другое слово). Таким образом, пространство признаков оказывается чрезвычайно большим, так что обработка этих репрезентации оказывается сильно затруднена; кроме того, многие из дистрибутивных признаков мало информативны и являются по сути дела шумом.

Проблема чрезвычайно большого числа измерений в пространстве признаков хорошо известна в области автоматической классификации документов, где документы подобным же образом представляются в виде многомерных признаковых векторов. Здесь эта проблема решаются с помощью вычисления веса для каждого признака, отражающего релевантность признака для отражения семантики документа. Взвешивание признаков (и/или селекция признаков на основе вычисленных весов) благоприятно сказывается как на эффективности классификации, так и на ее скорости (см., например, Yang&Pedersen, 1997). Цель настоящей работы – изучить перспективы применения методов взвешивания признаков, разработанных в области классификации документов, для автоматической классификации слов.

2. Функции взвешивания признаков

Общая процедура взвешивания признаков может быть описана следующим образом. Вес признака вычисляется из данных тренировочной части корпуса. Для каждого признака f_i из пространства признаков F тренировочной части взвешивающая функция (ВФ) вычисляет вес $w(f_i, c_j)$, локальный для каждого тренировочного класса $c_j \in C$. Классификатор (т.е., программа, приписывающая тестовые слова к классам) затем выводится из взвешенных

тренировочных репрезентаций и тестируется на тестовых данных. Во время классификации тестового слова, его репрезентация взвешивается весами, вычисленными из тренировочных данных о том классе, с которым тестовое слово сравнивается. Из множества весов признака, локальных для каждого класса, можно определить его единый глобальный вес с помощью

одного из методов глобализации: (1) суммирования локальных весов $w_{glob}(f_i) = \sum_{j=1}^{C_1} w(f_i, c_j)$; (2) вычисления их взвешенного среднего $w_{glob}(f_i) = \sum_{j=1}^{C_1} P(c_j) w(f_i, c_j)$; (3) определения максимального локального веса $w_{glob}(f_i) = \max_{j=1}^{C_1} w(f_i, c_j)$.

В литературе по классификации документов выделяются две стратегии, которые могут лежать в основе ВФ. Первая стратегия основывается на положении, что больший вес должен приписываться тем признакам, которые наиболее дифференцировано распределены между положительными и отрицательными примерами класса. Другими словами, чем чаще признак появляется у членов одного класса и реже у членов других классов, тем более его релевантность в характеристике семантики этого класса. Вторая стратегия основана на предположении, что больший вес должен сообщаться тем признакам, которые присутствуют у семантически схожих слов (например, у слов, принадлежащих одному семантическому классу).

В настоящей работе изучаются ранее не изученные варианты “силы признака”, ВФ, основанной на второй стратегии, и вводится новая ВФ, в основе которой лежит та же стратегия. В исследование включена также широко используемая при классификации документов ВФ “взаимная информация”, основанная на первой стратегии.

Взаимная информация (mutual information). Взаимная информация является заимствованной из теории информации и широко используемой в статистической ОЕЯ мерой ассоциации между двумя словами или какими-либо другими языковыми явлениями (Manning&Schuetze, 1999). Взаимная информация описывает количество информации, которое сообщает появление одного из двух слов о появлении другого. Она определяется как логарифм отношения эмпирически оцененной вероятности одновременного появления обоих слов к произведению вероятностей появления каждого из слов, взятого в отдельности:

$$MI(f_i, c_j) = \log \frac{P(f_i, c_j)}{P(f_i)P(c_j)} \quad (1)$$

Взаимная информация между признаком и классом описывает степень ассоциации признака и класса.

Сила признака (term strength). “Сила признака” как ВФ, применяемая для классификации документов, предложена в работах Й.Янг и ее коллег (см., например, Yang&Pedersen, 1997). Эта функция описывает вероятность появления признака в одной репрезентации, при условии, что этот признак присутствует в репрезентации, наиболее схожей с первой. В ее основе лежит мысль о том, что наиболее информативными являются признаки, являющиеся общими для наиболее схожих репрезентаций. Для вычисления силы признака сначала определяются пары наиболее семантически близких репрезентаций с помощью некой метрики дистрибутивной схожести, как, например, косинус между векторами признаков. Сила признака затем определяется как условная вероятность появления признака f_i в репрезентации R , если f_i присутствует и в семантически близкой R' :

$$S(f_i) = P(f_i \in R | f_i \in R') \quad (2)$$

Сила признака не использует данные о распределении тренировочных слов по классам и поэтому эта ВФ может быть только глобальной.

Мы предлагаем вариант силы признака, который в качестве ингредиента веса использует данные о распределении слов по классам: при формировании пар наиболее схожих репрезентаций, для каждого тренировочного слова наиболее схожее определяется не среди всех тренировочных слов, а только тех, которые принадлежат к тому же классу:

$$S_{class-based}(f_i) = P(f_i \in R_c | f_i \in R_c') \quad (3)$$

Таким образом, эта функция сообщает больший вес признакам, релевантным для дифференциации именно рассматриваемых классов слов.

Другой вариант силы признака, анализируемый в этой работе, является локальным – вес определяется локально для каждого класса на основе распределения признака по парам слов, сформированным из членов этого класса:

$$S_{local}(f_i, c_j) = P(f_i \in R_{c_j} | f_i \in R_{c_j}') \quad (4)$$

Используя эту функцию, мы надеемся еще больше подчеркнуть различия между рассматриваемыми классами слов.

Наконец, мы вводим новую ВФ, **типичность признака**, в основе которой лежит та же стратегия, что и для силы признака. В отличие от последней, в качестве схожих слов используются не пары наиболее схожих друг с другом слов, а все индивидуальные слова класса. Предположение, лежащее в основе этой ВФ, заключается в том, что признаками, наиболее релевантными для принадлежности классу, являются те, которые присутствуют у наибольшего числа членов этого класса. Типичность признака определяется как отношение числа членов класса, имеющих этот признак $n(f_i, c_j)$, к общему числу его членов $N(c_j)$:

$$TYP(f_i, c_j) = \frac{n(f_i, c_j)}{N(c_j)} \quad (5)$$

То, что вес признака зависит не от попарного объединения наиболее схожих слов, а от более крупных семантических групп слов, позволяет учесть закономерности объединения множества слов в эти группы.

В таблице 1 в качестве примера приводятся 10 дистрибутивных признаков для слова *car*, получивших наибольший вес согласно трем ВФ: взаимной информации, силе признака и типичности признака. В качестве признаков использовались глаголы, при которых *car* является прямым или предложным дополнением (веса были вычислены на основе данных Британского Национального Корпуса, см. раздел 3).

MI _{glob}		S		TYP	
admire	7.875	travel_in	0.75	drive	0.857
trap_in	7.789	get	0.716	see	0.857
pull	7.75	see	0.589	take	0.714
polish	7.551	make	0.569	get	0.714
free_from	7.363	take	0.532	get_into	0.571
jerk	7.347	put	0.526	want	0.428
slam	7.232	receive	0.521	go_by	0.428
leave_of	7.217	use	0.516	call	0.428

ride_in	6.993	drive_in	0.5	park	0.428
pick	6.814	escape_in	0.5	run_to	0.428

Таблица 1. 10 наиболее релевантных признаков слова *car* определенных с помощью трех ВФ: взаимная информация, сила признака, типичность признака.

Приведенные в таблице 1 примеры сообщаемых различными ВФ весов наглядно иллюстрируют коренные различия между двумя стратегиями взвешивания признаков. Первая стратегия, придающая больший вес признакам, уникальным для класса, воплощена в ВФ взаимная информация. Согласно этой ВФ, наибольший вес получают довольно редкие признаки, многие из которых весьма информативны (*ride_in*), но и многие же скорее являются окказионально встречающимися (*trap_in*) или ошибками синтаксического парсера (*leave_of*, *slam*). Вторая стратегия, сообщающая больший вес признакам, повторяющимся у семантически схожих слов, “предпочитает” высокочастотные признаки. Среди них есть как достаточно информативные (*drive*, *go_by*, *travel_in*), так и слишком абстрактные (*take*, *see*, *get*, *make*).

В следующих разделах будут изложены результаты количественной оценки изучаемых функций и их свойств.

3. Данные экспериментов

Классифицирующий алгоритм. В качестве классифицирующего алгоритма использовался метод ближайших соседей (KNN), который успешно используется в различных областях машинного обучения, включая классификацию документов (Yang&Pedersen, 1997) и дистрибутивную классификацию слов (Tokunaga et al. 1997; Pekar&Staab, 2002). При использовании KNN принадлежность нового объекта к некому классу определяется схожестью этого объекта с одним или более отдельными членами этого класса. Схожесть между объектами определяется с помощью метрики схожести, например, косинусом между признаковыми векторами.

Чтобы классифицировать новое слово, необходимо сначала определить некоторую группу наиболее схожих с ним слов тренировочной части, т.н. «ближайших соседей». Число слов в этой группе – важный параметр в этом алгоритме, который может варьироваться в зависимости от конкретной условий задачи, для решения которой применяется алгоритм. Новое слово приписывается к тому классу, который имеет наибольшее число своих членов в группе «ближайших соседей». Решение о классовой принадлежности нового слова может быть дополнительно основано на мере схожести между новым словом и его ближайшими соседями: «голос» каждого соседа за свой класс взвешивается пропорционально мере его дистрибутивной схожести с новым словом. В настоящем исследовании использовался “взвешенный” вариант KNN.

Дистрибутивная репрезентация. Конкретная задача, с помощью которой мы тестируем изучаемые ВФ, заключается в классификации существительных по семантическим классам. Семантика каждого существительного n представляется в виде вектора, где признаками являются глаголы $v \in V$, при которых n является прямым или предложным дополнением. Значения признаков – условные вероятности $P(v|n)$, вычисленные из наблюдаемых в корпусе частот.

Метрика дистрибутивной схожести. Для измерения схожести между векторами существительных n и m , использовалась метрика L_1 (также известная как city-block distance):

$$L_1(n, m) = \sum_{v \in V} |P(v_i | n) - P(v_i | m)| \quad (6)$$

Данные корпуса. В исследовании использовались две группы данных: данные о дистрибуции существительных из (1) Британского Национального Корпуса (БНК) и из (2) газетного корпуса “Associated Press 1988” (АП). Данные БНК составили 1,34 миллиона пар глагол-существительное; пары, которые встретились лишь однажды были удалены; были удалены существительные, которые встречались с менее чем пятью глаголами. Данные АП составили 0,73 миллиона пар; рассматривались только 1000 наиболее употребимых существительных этого корпуса.

Извлеченные таким образом из корпусов существительные были сгруппированы в семантические классы следующим образом. Для этой цели использовался электронный тезаурус WordNet: каждая синонимическая группа WordNet, располагающаяся на семь уровней ниже корневого уровня тезауруса, объединяясь со всеми своими гипонимами, образовывала отдельный класс существительных. Таким образом, из данных БНК было создано 60 классов с 514 существительными, а из данных АП – 43 класса с 375 существительными. Среднее число признаков в репрезентации существительного для группы данных БНК – 48,313, для группы данных АП – 162,89.

Метод тестирования. Для оценки качества классификации использовалось десятикратное перекрестное тестирование. Все существительные одного корпуса сначала делились на тренировочную и тестовую части в пропорции 9 к 1. После того, как классифицирующий алгоритм строил классификатор на основе данных тренировочной части, классификатор тестировался на данных тестовой части: проверялась его способность верно определить класс тестового существительного. Затем корпус по-новому делился на тренировочную и тестовую части так, что в качестве тестовых использовались новые существительные. Эта процедура повторялась десять раз, после чего вычислялось среднее мер эффективности, полученных при тестировании каждой из десяти тестовых частей.

Мера эффективности. Мерой эффективности служила точность классификации, определяемая как отношение числа верно классифицированных тестовых существительных к общему числу тестовых существительных.

4. Результаты экспериментов

В качестве эталона, с которым сравнивалось качество классификации, осуществляемых с помощью различных ВФ, использовались результаты, полученные для невзвешенных дистрибутивных репрезентаций.

4.1. Локальные и глобальные варианты ВФ

Сначала были проведены эксперименты по сравнению локальных и глобальных вариантов М1 и ТУР.

При оценке локальных вариантов мы столкнулись с проблемой, ранее описанной в (Debole&Sebastiani, in press), по причине которой локальный вариант функции, вопреки ожиданиям (вес, локальный для класса, со сути должен лучше дискриминировать этот класс от других, чем глобальный вес), оказывается значительно хуже ее глобальных вариантов. Эта проблема заключается в том, что во время классификации тестовые репрезентации взвешиваются весами тех тренировочных классов, с членами которых она сравнивается. При этом все признаки в тестовой репрезентации, отсутствующие в тренировочном классе получают вес, равный нулю. Поэтому, если между тестовой репрезентацией и тренировочным классом мало общих признаков, то тестовая репрезентация оказывается слишком разряженной и таким образом неспособной должным образом отразить семантику слова.

Чтобы решить эту проблему, мы сначала пробовали несколько способов сгладить локально взвешенные тестовые репрезентации:

1) Для вычисления нового значения признака f_i к его локальному весу $w(f_i, c_j)$ прибавлялся наименьший вес, имеющийся среди всех признаков F тренировочного класса c_j .

$$value(f_i) = value(f_i)(w(f_i, c_j) + \min_{k=1}^{|F|} w(f_k, c_j))$$

2) К значению взвешенного признака f_i прибавлялся его невзвешенное значение:

$$value(f_i) = value(f_j)w(f_i, c_j) + value(f_i)$$

3) В случае, если признак тестового слова отсутствует среди признаков тренировочного класса, использовался глобальный вес этого признака. Отметим, что при сравнении различных способов глобализации нами было обнаружено, что самым эффективным является использование максимального локального веса в качестве глобального. В дальнейшем, при обсуждении глобального веса ВФ речь всегда будет идти о этом способе глобализации.

4) Тестовые репрезентации вовсе не взвешивались.

В таблице 2 приводятся результаты оценки этих способов сглаживания (приводятся результаты для того числа ближайших соседей, для которого была получена наибольшая эффективность).

Способ сглаживания	БНК			АП		
	MI	S_{local}	TYP	MI	S_{local}	TYP
Без сглаживания	0,28823	0,14607	0,19656	0,27219	0,28506	0,35689
Минимальный вес	0,37409	0,23103	0,24721	0,39431	0,32522	0,37027
Невзвешенное значение	0,3331	0,34276	0,34852	0,34928	0,39744	0,39203
Глобальный вес	0,40746	0,30395	0,30554	0,38101	0,31566	0,40519
Невзвешенная репрезентация	0,37213	0,36187	0,39736	0,39431	0,39928	0,42852

Таблица 2. Способы сглаживания локально взвешенных тестовых репрезентаций.

Как видно этой таблицы, наибольшее качество классификации достигается в случае, если тестовая репрезентация не взвешивается вовсе: по сравнению с взвешенными, несглаженными репрезентациями этот способ имеет преимущество в ок.10% для MI и S_{local} (БНК и АП) и для TYP ок.20% на данных БНК и ок.7% на данных АП. Существенная разница между эффективностью TYP для групп данных БНК и АП для несглаженных репрезентаций по-видимому говорит о том, что эффективность TYP сильно зависит от числа признаков в дистрибутивной репрезентации (вспомним, что среднее число признаков в репрезентации для БНК значительно меньше этого среднего для АП; при локальном же взвешивании число признаков, как было отмечено выше, становится еще меньше).

Другим эффективным способом сглаживания является использование глобального веса для тех признаков тестовой репрезентации, которые отсутствуют в тренировочном классе. Эффективность этого способа является наивысшей для MI , данные БНК; этот метод второй по эффективности для данных АП как для MI так и для TYP .

Таким образом, в дальнейшем в качестве локальных вариантов ВФ, мы будем использовать те, при которых тестовая репрезентация не взвешивается.

После этого мы сравнивали локальные варианты этих ВФ с вычисленными из них глобальными вариантами. Результаты этих экспериментов приводятся в таблице 3.

	БНК			АП		
	MI	S_{local}	TYP	MI	S_{local}	TYP
Локальный	0,37213	0,36187	0,39736	0,39431	0,39928	0,42852
Глобальный	0,41874	0,31384	0,34856	0,40746	0,36904	0,40526

Данные обоих корпусов свидетельствуют о том, что для MI глобальный вариант является предпочтительным, в то время как для S_{local} и TYP – локальные.

4.2. Варианты силы признака

Таблица 3 описывает результаты сравнения различных вариантов силы признака (см. формулы 2, 3 и 4; приводятся результаты для того числа ближайших соседей, для которого была получена наибольшая эффективность).

	БНК	АП
S	0,311425	0,3867
$S_{class-based}$	0,323746	0,378663
S_{local}	0,361878	0,399289

Таблица 3. Различные варианты силы признака.

Как видно из этих результатов, для обоих групп данных наилучшим вариантом силы признака является ее локальный вариант.

4.3. MI, S, TYP

Наконец, мы сравнивали лучшие варианты трех изучаемых ВФ: MI_{glob} , S_{local} , TYP с эталоном (baseline) – классификатором, полученным из невзвешенных репрезентаций. Результаты этих экспериментов на данных БНК приводятся на рис.1, где показана эффективность ВФ как функция от числа ближайших соседей.

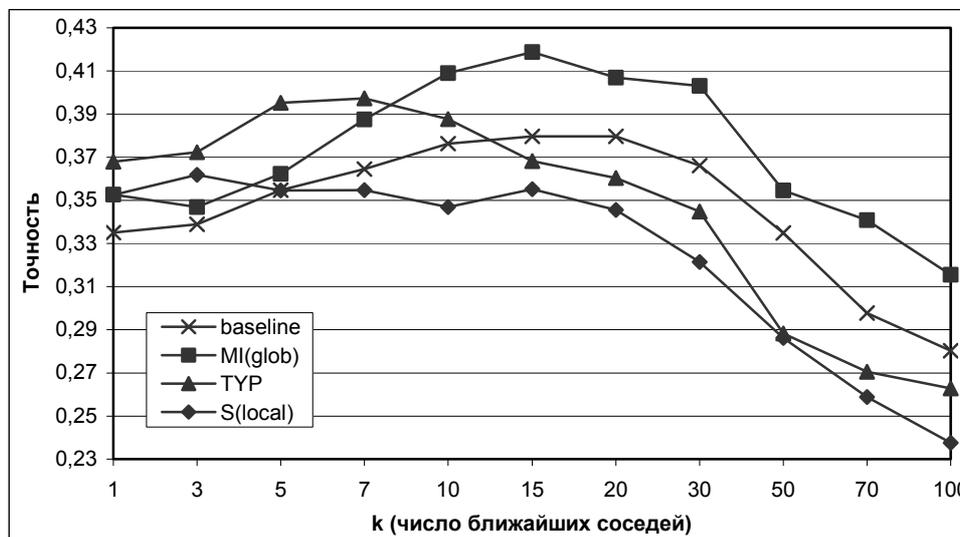


Рис. 1. Эффективность MI, TYP и S как функция от числа ближайших соседей (БНК).

Как видно из рис.1, лучшей ВФ для данных БНК является MI_{glob} (относительное преимущество над эталоном – 10,2%). TYP также имеет преимущество над эталоном – 4,6% и, кроме того, показывает наибольшую среди всех ВФ эффективность для $k < 7$. S_{local} , однако, имеет худшее качество классификации, чем эталон.

Рис 2. описывает результаты тестирования этих же ВФ на данных АП.

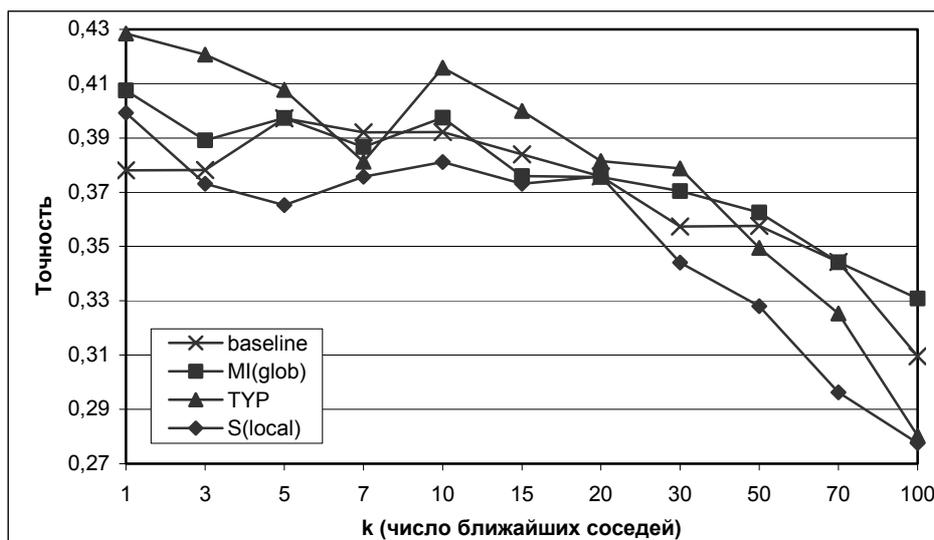


Рис. 2. Эффективность MI, TYP и S как функция от числа ближайших соседей (АП).

Для данных АП картина несколько иная. Наилучшей ВФ здесь является TYP (относительное преимущество над эталоном – 7,8%). Как и на данных БНК, его преимущество особо очевидно для малых значений k . Преимущество MI_{glob} над эталоном не столь велико как для данных БНК – 2,5%. S_{local} демонстрирует ту же эффективность, что и эталон.

4.4. Свойства функций

Учитывая несколько различные результаты, которые были получены для двух групп данных, мы провели эксперименты, направленные на выявление двух свойств MI_{glob} и TYP: зависимости их эффективности от (1) объема дистрибутивных данных, содержащихся в корпусе и (2) числа членов в тренировочном классе. Для этой цели были созданы еще две группы данных. Первая была сформирована из десяти классов БНК, имеющих наибольшее число членов. Вторая подобным образом была создана из данных АП. Таким образом, в этих экспериментах использовались четыре группы данных – БНК-60, АП-43, БНК-10 и АП-10. Зависимость ВФ от объема дистрибутивных данных изучалась путем сравнения результатов полученных, с одной стороны, для БНК-60 и БНК-10 и, с другой стороны, для АП-43 и АП-10. Зависимость ВФ от числа членов в тренировочном классе изучалась путем сравнения результатов полученных, с одной стороны, для БНК-60 и АП-43 и, с другой стороны, для БНК-10 и АП-10.

В таблице 4 приводятся результаты этих экспериментов (P – точность классификации, RI – относительное преимущество над эталоном).

	БНК-60		АП-43		БНК-10		АП-10	
	P	RI, %	P	RI, %	P	RI, %	P	RI, %
baseline	0,37967	0	0,39729	0	0,64921	0	0,594	0
MI_{glob}	0,41874	+ 10,2	0,40746	+ 2,5	0,67973	+ 4,7	0,6	+ 0,9
TYP	0,39736	+ 4,6	0,42852	+ 7,8	0,65368	+ 0,6	0,588	- 0,9

Таблица 4. Результаты оценки MI_{glob} и TYP на материале четырех групп данных.

Сравнивая результаты для БНК-60 и АП-43 с результатами БНК-10 и АП-10, мы видим, что качество классификаций ухудшается с увеличением числа членов в классе. Причина этого ухудшения может заключаться в том, что по мере увеличения числа членов в

классе, класс становится более гетерогенным, т.е. его члены сильно различаются по своей семантике, из-за чего оказывается сложно определить признаки, которые бы могли достаточно успешно дифференцировать этот класс от других.

При сравнении результатов для БНК-60 и БНК-10 с результатами АП-43 и АП-10, мы видим, что MI_{glob} гораздо эффективнее для данных БНК, т.е. для репрезентаций с меньшим числом признаков. ТУР же демонстрирует противоположную характеристику – с увеличением числа признаков в репрезентации улучшается и его эффективность (ср. 4,6% для БНК-60 с 7,8% для АП-43). Это его качество, однако, неочевидно из сравнения БНК-10 и АП-10, где эта функция имеет примерно одинаковую эффективность.

5. Заключение

В статье были изложены результаты исследования функций взвешивания признаков в применении к дистрибутивным репрезентациям слов. Сравнивались функции, воплощающие в себе две различные стратегии взвешивания признаков – взаимная информация и сила признака. Была также предложена новая функция – типичность признака, основанная, как и сила признака, на предположении, что наиболее релевантными при определении принадлежности классу являются те признаки, которые присутствуют у наибольшего числа членов этого класса.

Результаты исследования можно обобщить следующим образом.

1) Наиболее предпочтительным способом применения локальных вариантов изученных функций является использование невзвешенных тестовых репрезентаций.

2) Для силы признака и типичности локальные варианты являются предпочтительней, в то время как для взаимной информации – глобальный.

3) Были предложены два новых варианта силы признака, из которых локальный вариант оказался более эффективным, также и в отношении к ее “стандартному” варианту.

4) Наиболее предпочтительное качество классификаций было продемонстрировано взаимной информацией и типичностью признака. Оба метода были более эффективны для групп данных, где тренировочные классы имеют меньшее число членов, что можно объяснить тем, что классы с большим числом членов гетерогенны и поэтому для них трудно определить признаки, способные должным образом отразить индивидуальную семантику класса. Взаимная информация была более эффективна для тех групп данных, где репрезентации содержали меньшее число признаков. Типичность, напротив, – для тех групп данных, где репрезентации содержали большее число признаков.

Литература

Alfonseca E. and Manandhar S. (2002). Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures. // Proceedings of EKAW'02, pp.1-7.

Caraballo S. A. (1999) Automatic construction of a hypernym-labeled noun hierarchy from text. // Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 120-126.

Debole F. and Sebastiani F. (in press). Supervised term weighting for automated text categorization.

Hearst M.A. (1992). Automatic acquisition of hyponyms from large text corpora. // Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France.

- Lin D. (1998) Automatic retrieval and clustering of similar words. // Proceedings of the COLING-ACL'98, pp. 768-773.
- Lin D. and Pantel P. (2002). Concept discovery from text. // Proceedings of COLING'02. Taipei.
- Manning C. and Schuetze H. (1999). Foundations of statistical natural language processing. MIT Press.
- Pekar V. and Staab S. (2002). Factoring the structure of a taxonomy into a semantic classification decision. // Proceedings of COLING-2002. Taipei. pp. 786-792.
- Roark B. and Charniak E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. // Proceedings of COLING-ACL'98. pp.1110-1116.
- Tokunaga T., Fujii A., Iwayama M., Sakurai N., and Tanaka H. (1997). Extending a thesaurus by classifying words. // Proceedings of the ACL-EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, pp. 16-21.
- Yang Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. // Proceedings of the 14th International Conference on Machine Learning. Nashville, TN, pp.412-420.