**Метод автоматизированного пополнения онтологии за счет анализа групп однородных прилагательных**

Клышинский Э.С., Кочеткова Н.А.

МИЭМ НИУ ВШЭ

**Введение**

На данный момент онтология является рабочим инструментом при создании большого количества информационных систем. В связи с этим активно разрабатываются методы, позволяющие автоматизировать процесс внесения информации в онтологии (см., например, [Рубашкин 2010]).

В данной статье рассматривается метод автоматизации процесса пополнения онтологии общей направленности, на основе автоматической кластеризации прилагательных в семантически связанные группы. Полученные кластеры в дальнейшем могут использоваться для группового внесения в онтологию с приписыванием им общих свойств.

**Метод получения исходного материала**

В качестве исходных данных для нашего метода были выбраны группы семантически связных прилагательных. Для их выделения было решено использовать группы однородных и неоднородных прилагательных. Заметим, что современные определения однородности и неоднородности не являются строго формальными и, следовательно, неудобны для точного толкования при разработке систем автоматической обработки текстов. Так, согласно [Валгина 2000] «Однородными членами предложения называются одноименные члены, связанные друг с другом сочинительной связью и выполняющие одинаковую синтаксическую функцию в предложении, т.е. объединенные одинаковыми отношениями к одному и тому же члену предложения». В основу нашего метода легло определение неоднородных членов, а именно определений (так как мы имели дело только с прилагательными), которое гласит: «Определения являются неоднородными, если предшествующее определение относится не непосредственно к определяемому существительному, а к сочетанию последующего определения и определяемого существительного…», и далее «Неоднородные определения характеризуют предмет с разных сторон, в разных отношениях…» [Валгина 2000]. В данном определении явно отражено то, что неявно подразумевается для однородных членов, а именно то, что однородные члены описывают различные значения одного и того же или сходных свойств объекта.

Предлагаемый нами метод выделения однородных членов основывается на результатах наших предыдущих работ [Клышинский 2011]. Из корпуса текстов, достаточно большого объема, извлекаются конструкции вида «прилагательное, прилагательное {, прилагательное}\* {и/или прилагательное} существительное». Здесь текст, заключенный в фигурные скобки, является необязательным, а знак \* означает, что текст может встречаться (если встречается) несколько раз. Для выделения неоднородных членов использовался шаблон «прилагательное прилагательное {прилагательное}\* существительное». Кроме того, как для однородных, так и для неоднородных определений, должен существовать хотя бы один вариант согласования всех прилагательных, составляющих группу с данным существительным. Для повышения точности рассматривались только группы неомонимичных по части речи слов, приведенных к нормальной форме.

Эксперименты выявили недостатки метода, относящиеся к корректности анализируемых текстов. Дело в том, что корректная расстановка запятых в подобных конструкциях регулируется скорее семантическими правилами, апеллирующими к здравому смыслу автора. В связи с этим часто возникают спорные ситуации, в которых делаются ошибки. Кроме того, группы однородных прилагательных указанного вида (с учетом неомонимичности всех слов) встречаются достаточно редко (одна на 5-10 страниц текста), то есть для того, чтобы получить базу большого объема необходимо использовать корпуса текстов действительно большого размера (от десятков миллионов слов).

Также в русском языке в ряде случаев запятой отделяются некоторые определения, не являющиеся однородными. Так, например, «Схожесть признаков может возникнуть у прилагательных, употребляемых в переносном значении: *Я пожал протянутую мне большую, черствую руку* (Шол.);…» [Лопатин 2009]. Также запятыми будут разделяться неоднородные определения, стоящие после определяемого слова (например, у А.П. Чехова «…я видел *женщину молодую, прекрасную, добрую, интеллигентную, обаятельную*, женщину, какой я раньше никогда не встречал…»). Таким образом, можно сделать вывод о том, что группы, полученные для литературных текстов будут априори содержать в себе определенный процент ошибок.

**Метод выделения связных групп прилагательных**

Как это отмечалось выше, однородные прилагательные внутри одной группы должны быть семантически связаны между собой. Далее необходимо решить вопрос связности прилагательных между выделенными группами. Для этого строится граф связности, вершинами которого служат все выделенные уникальные прилагательные. Дуги соединяют между собой прилагательные, которые встречаются в одной группе.

После создания графа проводится фильтрация связей, которые могли быть включены в граф по ошибке. Для этого анализируются группы неоднородных прилагательных. Пусть в графе есть связь между некоторыми двумя прилагательными. Если существует такая группа неоднородных прилагательных, в которой представлены эти два связанных прилагательных, то связь между ними в графе разрывается.

В результате отсеивания получается несвязный граф, имеющий по результатам экспериментов около 10-20% вершин, не связанных с другими вершинами. Кроме того, оставшаяся часть графа оказывается разделенной на несколько групп связности, причем в одну из них включается значительное большинство вершин, тогда как в других может содержаться всего несколько вершин. Большая группа связности образуется по различным причинам. Во-первых, это наличие ошибок в выделенных группах, приводящих к паразитным связям. Во-вторых, многие слова являются полисемичными и связывают собой несколько кластеров. Хорошее отделение кластера на первом этапе возможно только при условии использования специальной лексики, образованной без использования заимствований, что встречается достаточно редко. Вспомним, например, характеристики кварков (странный, прелестный и др.), взятые из характеристик человека.

В связи с наличием большого кластера встает задача дальнейшего его разделения на подкластеры. Заметим, что дальнейшие рассуждения без нарушения общности можно перенести и на случай нескольких больших кластеров. В связи с тем, что данные кластеры будут несвязными, их деление на подкластеры может проводиться независимо, а результаты разделения могут быть объединены.

**Метод кластеризации графа связей между однородными прилагательными**

Применение общепринятых методов кластеризации графов в данном случае не дает приемлемого результата. Так, например, веса связей не присвоены, в связи с чем по данному признаку нельзя отбросить часть из них. Алгоритм Краскала [Кормен 2006] и сходные с ним выделяют в итоге единый кластер, включающий в себя все вершины графа. Так же ведут себя и многие агломеративные алгоритмы по причине наличия в графе «шума».

В ходе исследования был предложен следующий метод. На первом этапе ведется расчет матрицы расстояний между вершинами. Для этого можно использовать один из общепринятых алгоритмов, например, Флойда-Уоршела [Кормен 2006]. Далее кластеризация проводится следующим образом. Вводится вектор принадлежности вершин кластерам. Данный вектор хранит номер кластера, которому принадлежит данная вершина, или информацию о том, что вершина еще не была включена ни в один кластер. Из вершин графа, еще не участвовавших в кластеризации, выбираются две, расстояние между которыми максимально. Если расстояние между вершинами меньше заданного порогового значения, работа алгоритма прекращается. В противном случае вокруг каждой из вершин образуется новый кластер, состоящий из вершин, соседних данной. В нашем случае брались все вершины, лежащие на расстоянии не больше двух от выбранной.

По окончании работы алгоритма кластеризации для всех вершин, не присоединенных ни к одному кластеру, рассчитывается количество связей длины 1 со всеми кластерами. Вершина присоединяется к кластеру, количество связей с которым максимально.

Далее проводится объединение кластеров. Это связано с тем, что вершины, выбранные в качестве центров кластеров, на самом деле могут лежать на его краю. Кроме того, диаметр кластера может быть больше 5. В обоих случаях полученный кластер не включает в себя всех нужных вершин и разбивается на несколько. Для устранения этого недостатка мы объединяли все кластеры, между которыми есть связи длины 1 и максимальная связь между вершинами двух кластеров не превышает 5. Полученное таким образом разбиение на кластеры считается окончательным.

**Результаты экспериментов**

Для оценки работы метода был взят корпус текстов на основе библиотеки Мошкова (более 600 млн слов) и корпус новостных текстов различной направленности (например, Лента.ру, РИА Новости, РБК, PCWeek и Компьюлента, общий объем более 400 млн слов). Было проведено два эксперимента: на текстах из библиотеки Мошкова и на всей коллекции, в ходе которых проводилась кластеризация групп прилагательных.

В ходе эксперимента с текстами из библиотеки Мошкова было выделено более 8100 групп однородных прилагательных и около 620 тыс. неоднородных. Группы однородных членов составили 4439 различных прилагательных, из которых около 1000 не образовали связей с другими прилагательными (точнее, связи были отсеяны с использованием информации о группах неоднородных членов), а около 500 прилагательных образовали кластеры, не имеющие связей с другими кластерами. Всего было получено около 330 кластеров.

В экспериментах над всей коллекцией было получено около 32000 групп однородных прилагательных и более 2,6 млн групп неоднородных прилагательных. В группы однородных прилагательных вошли 6815 различных прилагательных. Из них примерно у 1250 не осталось каких-либо связей с другими словами, а примерно 450 прилагательных образовали компактные кластеры по 3-5 слов, которые не имели ни одной связи с другими кластерами. Всего было выделено около 420 кластеров.

 Качество кластеризации в обоих случаях было сходным. С одной стороны, хорошо выделились, например, кластеры, относящиеся к животным, растениям или географии. С другой стороны, географических кластеров было выделено несколько. При этом различные географические кластеры различались как по смыслу (города против стран и местностей), так и по географическому признаку (европейские города были отделены от российских). Однако и российские топонимы образовали несколько кластеров, то есть формально в кластеризации были допущены ошибки.

Среди кластеров оказалось около 80% таких, которые содержали ошибки различного рода. Во-первых, это ошибки, связанные с неверным отнесением слова к кластеру, когда в однородном по составу кластере есть одно или несколько слов, не связанных с ним по смыслу или семантически близких, но не относящихся к данной группе (например, *норковый, моржовый, тюлений, овечий, бычий, козий, поросячий, свиной, крысиный, кошачий, собачий, \*звериный, \*инстинктивный*). Кроме того, наличие подобной связи могло приводить к ошибочному объединению целых кластеров. И, наконец, полисемичность слов приводила в ряде случаев к объединению в один кластер нескольких разнородных слов (например, *батраческий, бедняцкий, кулацкий, середняцкий, пролетарский, краснопресненский, октябрьский, декабрьский, ноябрьский*).

Количество кластеров, объединяющих слова, относящиеся к различным свойствам, приблизительно оценивается в 20% от их общего числа. Количество кластеров, не содержащих ни одного или только одно ошибочное слово, оценивается также примерно в 20%.

**Выводы и перспективы развития**

Исходя из анализа полученных данных, можно сказать, что разработанный метод пригоден для автоматизации процессов заполнения онтологий общей направленности. Метод позволяет выделить смысловые группы прилагательных, которые далее могут быть откорректированы вручную. Однако кластеры могут рассматриваться с различных точек зрения. Так, например, кластер, содержащий в себе следующий фрагмент: *ивовый, миртовый, пальмовый, апельсиновый, лавровый, оливковый, сосновый, грейпфрутовый, банановый, хвойный, березовый*;является правильным с точки зрения свойства «иметь отношение к деревьям». Однако наличие слова «банановый» несколько размывает это свойство, так как формально банан является травой. Если рассматривать слова «апельсиновый» и «грейпфрутовый» с точки зрения свойства «виды сока», то кластер следует считать ошибочным. То есть при создании специализированных онтологий метод будет выдавать слишком общие результаты. Кроме того, для получения результатов метод требует больших текстовых баз, которые не всегда можно найти в случае узкой тематики. Вообще, сходство однородных определений трактуется достаточно широко, в связи с чем подобный подход слишком широк для автоматического применения в системах онтологий, хотя и пригоден для получения первого приближения.

Предложенный метод нуждается в ряде доработок. Самой важной из них является реализация алгоритма нечеткой кластеризации с тем, чтобы одно слово могло относиться к нескольким кластерам. Также нам представляется, что использование статистических методов снятия омонимии должно существенно расширить количество извлекаемых конструкций, однако внесет некоторый шум. Как показала практика, даже небольшой процент шума может привести к значительному снижению качества результатов. Применение же эмпирических правил (например, существительное в конце группы может быть омонимично глаголу) должно повысить количество выделяемых конструкций без нанесения урона качеству анализа.

В перспективе планируется повысить качество кластеризации за счет информации о совместной встречаемости прилагательных.

**Список литературы**

[Валгина 2000] Валгина Н.С. Синтаксис современного русского языка: Учебник / М.: Агар, 2000.  416 с.

[Клышинский 2011] Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Метод разрешения частеречной омонимии на основе применения корпуса синтаксической сочетаемости слов в русском языке // Научно-техническая информация. Сер. 2: Информационные системы и процессы. Серия 2. №1 2011 г., сс. 31-35

[Кормен 2006] Кормен Т.Х., Лейзерсон Ч.И., Ривест Р.Л., Штайн К. Алгоритмы: построение и анализ – М.: «Вильямс», 2006. – 1296 c.

[Лопатин 2009] Правила русской орфографии и пунктуации. Полный академический справочник / Под ред. В.В. Лопатина - М: АСТ, 2009. - 432 с.

[Рубашкин 2010] Рубашкин В.Ш., Бочаров В.В., Пивоварова Л.М., Чуприн Б.Ю. Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). - М.: Изд-во РГГУ, 2010. - сс. 413-418

**The automated method of ontology constructing
using information about homogeneous adjectives**

Klyshinsky E.S., Kochetkova N.A.

MIEM HSE

**Keywords:** automated ontology constructing, homogeneous adjectives, clustering, semantic groups

**Abstract:** The article describes a method that uses a clustering method for semantic grouping of adjectives. The method extracts semantically tailed adjective groups from untagged texts. These groups can be used for automated ontology constructing.