**<http://www.intuit.ru/studies/courses/1078/270/lecture/6861?page=2>**

[**http://www.myshared.ru/slide/276976/**](http://www.myshared.ru/slide/276976/)

**8.4. WordNet: проблемы использования в автоматической обработке текстов**

**Смешение типов и ролей**

Одной из серьезных проблем, приводящих к неправильным путям иерархии, является проблема установления таких отношений, когда вышестоящее понятие частично характеризует нижестоящее. Часто это связано с проблемой смешения понятий-типов и понятий-ролей.

Так, например, Никола Гуарино критикует отношения в WordNet: "Человек всегда живое существо, но он (она) начинает играть роль каузального агента только в некоторых ситуациях. Та же проблема возникает для яблока, которое всегда плод растения, а в некоторых ситуациях может быть пищей. Проблема в том, что человек и яблоко - это типы сущностей, в то время как каузальный агент и пища - это роли".

Один из аргументов в пользу различения типов и ролей в *лингвистических онтологиях* - это то, что они различаются в способах наследования свойств. WordNet не различает эти два типа понятий и помещает их в одни и те же иерархии.

В соответствии с онтологическими подходами понятия-типы не должны находиться в иерархиях ниже понятий-ролей. Более радикальный подход заключается в том, чтобы разделить *иерархии типов* и ролей.

Одна из авторов WordNet, Кристиан Фелбаум, отвечая на эту критику Н. Гуарино, заявляет, что в таких ресурсах, как WordNet, неоднородные классификации имеют право на существование, поскольку такие ресурсы рассматриваются в настоящее время прежде всего как инструменты для компьютерной обработки текстов, а не только как совершенные онтологии, которые должны соответствовать строгим онтологическим принципам.

Вместе с тем важно подчеркнуть, что описание понятий-типов ниже понятий-ролей, то есть установление связей между понятиями, которые выполняются не в любых контекстах, а лишь при некоторых условиях, приводит к ложному срабатыванию этих связей, к неправильному выводу как раз при автоматической обработке текстов.

**Многозначность в WordNet**

Во многих работах признается, что различия значений в WordNet слишком тонки для таких компьютерных приложений, как *машинный перевод*, информационный поиск, классификация текстов, вопросно-ответные системы и др. Среднее количество значений в WordNet больше, чем в традиционных лексикографических словарях.

Поэтому выполнен ряд исследований, чтобы разобраться, нельзя ли автоматически, на основе каких-либо разумных принципов собрать некоторые значения многозначных слов в кластеры и далее не различать эти значения.

**Проблема лексической многозначности и информационный поиск**

В частности, исследовался вопрос, какая группировка значений была бы полезной для задач информационного поиска. Предполагается, что некоторые значения могут быть кластеризованы для разных приложений; в то же время существуют примеры пар значений, кластеризация которых была бы полезна в информационно-поисковых приложениях, при этом в других приложениях было бы полезно их различать.

Отмечается, что исследования регулярной многозначности не приводят к выделению полезных кластеров для информационно-поисковых задач, поскольку, как представляется авторам, некоторые образцы регулярной полисемии хорошо бы не различать для задач информационного поиска, в то время как другие хорошо бы сохранить отдельно. Так, например, полезно было бы кластеризовать такие пары регулярной полисемии, как container / *quantity* и music / dance. Однако такие образцы, как animal / food, plant /food, animal / skin, language / people хорошо бы различать, поскольку, как представляется, они употребляются в разных типах текстов.

Поэтому нужны дополнительные исследования критериев кластеризации значений для информационно-поисковых задач.

Далее сравниваются два дополнительных критерия группировки значений. Первый критерий заключается в том, чтобы группировать значения, которые встречаются в одних и тех же текстах. Для этого используется семантически размеченный корпус SemCor. Второй критерий группирует значения, которые получают одни и те же переводы в нескольких разных языках. Пересечение кластеров, построенных на основе этих двух критериев, составляет 55-60%, что показывает некоторую корреляцию между кластерами, но оставляет сомнения в полезности каждого из критериев.

Проведенные эксперименты по кластеризации значений привели к выводу, что типология отношения между разными значениями многозначных слов является более полезной, чем формирование кластеров значений, поскольку близость значений зависит от приложения.

Например, указание, что одно из значений является метафорой исходного значения, является важным для приложений информационного поиска и вопросно-ответных систем, поскольку относится к разным семантическим полям. Однако для приложений *машинного перевода* это различие может быть несущественно, поскольку метафорический перенос может быть сходным в разных языках. В ворднетах нужно в явном виде описать отношения между значениями для того, чтобы ворднеты стали стандартом лексических ресурсов для компьютерных приложений.

Современные версии WordNet содержат для каждого многозначного слова указание на самое частотное значение по корпусу SemCor, что дает возможность, в случае проблем при процедуре автоматического разрешения многозначности, выбирать это самое частотное значение.

**Теннисная проблема**

Одной из серьезных проблем WordNet, препятствующей его использованию в приложениях, является так называемая "теннисная проблема":*синсеты*, принадлежащие одной предметной области, сфере деятельности или ситуации - оказываются очень далеко друг от друга в структуре WordNet.

Предлагалось решать данную проблему введением в WordNet информации о принадлежности синсетов определенным доменам. Домены, такие как "теннис", "политика" или "образование", группируют *синсеты* в сценарии или схемы. Так, домен "теннис" включает такие *синсеты*, как "гейм", "теннисный мяч", "теннисная ракетка", "тайм-брейк" и т.д.

Предполагается, что введение доменов должно быть особенно полезно для информационно-поисковых задач.

Разработка иерархической системы доменов началась с 250 рубрик, собранных по различным словарям, и затем была дополнена и уточнена на базе Десятичной классификации Дьюи. Была получена иерархия из 115 доменов, организованных по 4 уровням иерархии, которая включала, например, такие домены, как "сельское хозяйство", "археология", "астрология", "биология", "ветеринария" и др.

Авторы подчеркивают, что в WordNet имеются *синсеты*, которые не принадлежат никаким доменам, поскольку могут употребляться в текстах многих предметных областей. Для таких синсетов была введена специальная предметная область, называемая FACTOTUM.

Для того чтобы разметить все множество синсетов WordNet, была реализована автоматизированная процедура, состоящая из следующих шагов:

1. вручную размечается относительно небольшое количество синсетов верхнего уровня;
2. автоматически по связям (гипонимия, тропонимия, меронимия, *антонимия*) пометки распространяются на другие *синсеты*;
3. можно задать исключения, например, для *синсета* " кресло парикмахера " (" barber\_chair "), которое хотя и является частьюпарикмахерской (" barbershop "), не должно быть отнесено к домену КОММЕРЦИЯ ( COMMERCE ).

Эксперименты с доменами в ворднетах были продолжены и в следующем европейском проекте, связанном с ворднетами, - Meaning, в котором 165 иерархически организованных доменов были автоматизированно приписаны всем *синсетам* WordNet.

Авторы также подчеркивают полезность разметки синсетов доменами для автоматического разрешения лексической многозначности.

Вместе с тем остаются вопросы по отношению к введению в систему, построенную на основе одних единиц, набора других единиц с неопределенным относительно исходных единиц статусом, среди которых:

* вариативность возможного набора областей;
* небольшая наполненность некоторых доменов, и большое количество синсетов в других доменах:
* необходимость разных систем доменов для разных задач;
* отсутствие полностью выверенного набора доменов (выверять вручную очень трудоемко, а если выверять в процессе решения различных задач, то далеко не все проблемы (неточности, ошибки, приписки) удастся быстро обнаружить).

**Представление толкований WordNet в виде логических выражений: проект eXtended WordNet**

Многие исследователи отмечают нехватку информации, описанной в WordNet, для различения значений, в нем перечисленных.

В рамках проекта eXtended WordNet разработчики предполагают, что важным источником дополнительной информации могут стать толкования, приписанные к *синсетам* WordNet. Для того чтобы эти толкования можно было использовать в автоматических режимах компьютерных приложений, необходимо каждому знаменательному слову толкований сопоставить его значение-синсет и представить это толкование в виде формализованного выражения.

**Разрешение лексической многозначности**

Поскольку стало ясно, что применению таких ресурсов, как WordNet, препятствует такая проблема, как недостаточная эффективность разрешения лексической многозначности, эта проблема получила отдельную значимость.

Была организована специальная конференция, посвященная проблеме разрешения лексической многозначности - Конференция SENSEVAL.

Первая конференция по оценке методов разрешения лексической многозначности SENSEVAL состоялась в 1998 году, охватывала три языка, в ней приняли участие 25 исследовательских групп. Вторая состоялась в 2001 году, имела задания на 12 языках; в ней участвовали 35 исследовательских групп и более 90 систем.

**Основные принципы разработки, создания и использования традиционных информационно-поисковых тезаурусов. Примеры тезаурусов**

Начало разработки *информационно-поисковых тезаурусов* для различных предметных областей относится к середине 60-х годов. В то время большинство информационных систем не являлись полнотекстовыми, а хранили достаточно ограниченный набор информации о документе: библиографические данные, реферат. Добавление списка ключевых слов, характеризующих основное содержание документа, существенно расширяло возможности поиска документов. С начала 70-х годов создаются национальные и международные стандарты разработки*информационно-поисковых тезаурусов*.

*Значимость* разработки и использования *информационно-поисковых тезаурусов* значительно снизилась с появлением полнотекстовых информационно-поисковых систем, а также возможностей поиска *по* всем словам текста с помощью методов ранжированного информационного поиска, поскольку такие системы обеспечивают возможность поиска информации неподготовленному пользователю в любых предметных областях, без предварительных затрат на разработку тезаурусов.

Потенциально использование тезаурусов в качестве средств для описания основного содержания текста позволяет преодолевать многие проблемы пословного поиска, упомянутые в лекции 3, а также проблему, связанную с избытком приписанных тексту слов, которых оказывается так много, что возникает отдельная задача *по* определению их значимости для данного текста.

Однако многочисленные исследования *по* определению эффективности различных методов представления документов при информационном поиске показали, что эффективность пословного индексирования сравнима с эффективностью поиска, использующего ручное индексирование *по* тезаурусу.

Действительно, применение хорошо разработанного тезауруса при ручном индексировании должно снимать проблемы синонимии, близких понятий, многозначности. Однако при этом могут возникнуть существенные различия между понятиями, используемыми в тезаурусе, и информационной потребностью пользователя, когда пользователю трудно сформулировать описание нужных ему текстов посредством понятий тезауруса или *тезаурус* действительно не содержит адекватных понятий. В этих случаях пословное индексирование имеет преимущество из-за больших выразительных возможностей.

Кроме того, при ручном индексировании серьезную проблему составляет фактор субъективности, когда приписывание тексту терминов тезауруса зависит от умения и опыта индексаторов, от количества текстов, которые необходимо проиндексировать, и т.п.

Тем не менее и в настоящее время существуют информационные службы, имеющие и разрабатывающие информационно-поисковые тезаурусы, а также имеющие штат профессиональных индексаторов, индексирующих документы на основе тезаурусов. Примерами таких организаций являются Исследовательская служба Конгресса США, индексирующая *по* тезаурусу LIV (Legislative Indexing *Vocabulary*); Организация *по* продовольствию и сельскому хозяйству при ООН (FAO - Food *and* Agriculture Organization), разрабатывающая *тезаурус*AGROVOC; службы Европейского сообщества, использующие для индексирования Европейского законодательства *тезаурус* EUROVOC и др. Происходит и процесс обновления стандартов разработки тезаурусов.

За прошедшие годы были разработаны и использовались информационными и терминологическими службами сотни тезаурусов, каждый из которых содержит ценную информацию о своей *предметной области*. Поэтому многие разработчики автоматических информационных систем исследовали вопросы о применении существующих *информационно-поисковых тезаурусов* при обработке документов в автоматическом режиме. Однако подавляющее большинство экспериментов окончились неудачей: применение *информационно-поисковых тезаурусов* в процессе автоматического индексирования увеличивало полноту поиска, но резко снижало его *точность*.

Более того, *международный стандарт* *по* разработке одноязычных тезаурусов (*ISO* 2788) четко указывает, что стандарт должен применяться в организациях, имеющих людей-индексаторов, которые анализируют содержание документов и описывают основные темы документов с помощью терминов тезауруса. "Применение стандарта не предполагает его применение в тех организациях, которые используют полностью автоматические методы индексирования".

Возникает вопрос: почему существующая *парадигма* разработки *информационно-поисковых тезаурусов* не дает возможности использовать созданные ресурсы в автоматических режимах индексирования текста? Можно ли и как именно создавать тезаурусы для автоматического индексирования? Для этого необходимо разобраться, какие особенности существующей парадигмы разработки *информационно-поисковых тезаурусов* не позволяют их использовать в автоматических режимах.

В дальнейшем тексте информационно-поисковые тезаурусы, создаваемые в соответствии с существующими международными и национальными стандартами, будем называть традиционными *информационно-поисковыми тезаурусами*.

**Назначение информационно-поисковых тезаурусов**

В различных стандартах и пособиях приводятся разные определения *информационно-поисковых тезаурусов*. Объемлющее *определение**информационно-поискового тезауруса* можно сформулировать следующим образом:

***Информационно-поисковый тезаурус*** (ИПТ) - это *контролируемый словарь* терминов на естественном языке, явно указывающий отношения между терминами и предназначенный для информационного поиска.

Основными целями разработки традиционных ИПТ являются следующие:

* обеспечение перевода естественного *языка документов* и пользователей на *контролируемый словарь*, применяемый для индексирования и поиска;
* обеспечение последовательного использования единиц индексирования;
* описание отношений между терминами;
* использование как поискового средства при поиске документов.

**Единицы традиционных информационно-поисковых тезаурусов**

Основной единицей тезаурусов являются *термины*, которые разделяются на *дескрипторы* (авторизованные термины) и недескрипторы (*аскрипторы* ).

Большинство версий стандартов *по* ИПТ указывают на *связь* терминов с *понятиями* *предметной области*. *По* американскому стандарту термин - это *слово* либо *словосочетание*, обозначающее понятие. Стандарт *ISO* подчеркивает, что индексирующий термин - это*представление* понятия предпочтительно в форме существительного или именной группы.

При этом понятие рассматривается как *единица* мысли, которая формируется мысленно для отражения всех или некоторых свойств конкретного или абстрактного, реально существующего или мысленного объекта. Понятия существуют как абстрактные сущности, независимо от терминов, которые их выражают.

Российский ГОСТ рассматривает понятие как форму мышления, отражающую существенные свойства, связи и отношения предметов и явлений, а термином в определении ГОСТа является *слово* или *словосочетание*, являющееся точным обозначением определенного понятия какой-либо области знания.

ГОСТ 7.74-96 определяет единицы тезауруса как лексические единицы информационно-поискового языка - то есть обозначения отдельного понятия, принятые в информационно-поисковом языке и неделимые в этой функции.

Стоит отметить, что не все разработчики тезаурусов четко разделяли понятия и термины. Так, разработчики тезауруса AGROVOC определили его как термино-ориентированный (term-oriented), что находит свое проявление в том, что к термину невозможно добавить синонимы. Эта особенность тезауруса рассматривается авторами как недостаток, который необходимо исправить.

Таким образом, разработчики тезаурусов предполагают, что понятие *предметной области* обычно имеет несколько возможных вариантов лексического представления в тексте, которые рассматриваются как синонимы. Среди таких синонимов выбирается ***дескриптор*** - термин, который рассматривается как основной способ ссылки на понятие в рамках тезауруса. Другие термины из синонимического ряда, включенные в тезаурус, называются ***аскрипторами*** или недескрипторами. Они используются как вспомогательные элементы, текстовые входы, помогающие найти подходящие дескрипторы.

**Дескрипторы**

Дескрипторы тезауруса должны соответствовать выбранной предметной области тезауруса. Каждый дескриптор, внесенный в тезаурус, должен представлять отдельное понятие данной области. Дескриптор может быть однословным или многословным. Поскольку часто бывает достаточно трудно понять, представляет ли отдельное понятие многословное словосочетание, многие тезаурусы и руководства уделяют особое внимание основным принципам включения в тезаурус в качестве дескрипторов многословных терминов.

Набор дескрипторов должен удовлетворять следующим требованиям:

* посредством выделенных дескрипторов должно быть возможно описать темы абсолютного большинства текстов предметной области;
* для уменьшения субъективности индексирования множество дескрипторов не должно включать совокупности близких дескрипторов; для этого формируются классы условной эквивалентности, когда совокупности близких, но различных понятий сводятся к одному дескриптору;
* дескриптор должен быть сформулирован однозначно, его подразумеваемое в рамках тезауруса значение должно быть понятно пользователю. Если однозначный и ясный дескриптор подобрать не удается, термин, взятый в качестве дескриптора, снабжается релятором (краткой пометкой) или комментарием.

**Отношения в информационно-поисковом тезаурусе**

ГОСТ 7.25 указывает, что основными типами отношений, обычно отражаемых в ИПТ, являются следующие:

* род - вид;
* часть - целое;
* причина - следствие;
* сырье - продукт;
* административная иерархия;
* процесс - объект;
* функциональное сходство;
* процесс - субъект;
* свойство - носитель свойства;
* *антонимия*.

Такие содержательные *типы связей* между дескрипторами чаще всего не отражаются в подробном перечне отношений тезауруса, а записываются с помощью небольшого набора отношений, которые обычно разделяются на два типа: *иерархические* и *ассоциативные*.

**Иерархические отношения**

По ГОСТУ 7.25-2001 *иерархические отношения* обладают свойствами *транзитивности* и *антисимметричности*, которые могут быть использованы при избыточном индексировании в интересах повышения эффективности информационного поиска. Предпочтительно указывать связи между дескрипторами как отношения иерархического вида, если они обладают этими свойствами. Применяемые в ИПТ*иерархические отношения* могут дифференцироваться на отдельные виды.

Основным иерархическим отношением, используемым в ИПТ, является родо-видовое отношение (оно же - отношение НИЖЕ-ВЫШЕ ). По ГОСТУ 7.25-2001 родо-видовая связь устанавливается между двумя дескрипторами, если объем понятия нижестоящего дескриптора входит в объем понятия вышестоящего дескриптора.

Также в качестве иерархического отношения в ИПТ может устанавливаться отношение ЧАСТЬ-ЦЕЛОЕ.

Многие руководства и стандарты подчеркивают, что *иерархические отношения* в ИПТ должны устанавливаться в тех случаях, когда отношения истинны независимо от контекста - только в таких случаях дескрипторы ИПТ могут быть организованы в иерархии. Эта рекомендация связана с тем, что обычно в информационном поиске очень трудно четко определить контекст употребления термина и понять, применимо ли в данном контексте то или иное отношение.

Так, для мышей можно указать, что они грызуны, поскольку это внутренняя характеристика мышей. В то же время неправильно указывать, что мыши - вредители, поскольку имеются лабораторные мыши и домашние мыши, которые вредителями не являются.

Рекомендуется использовать тест "все-некоторые". Например, "все мыши являются грызунами, но некоторые мыши являются вредителями".

**Ассоциативные отношения**

Основное назначение установления ассоциативных отношений между дескрипторами ИПТ - указание на дополнительные дескрипторы, полезные при индексировании или поиске.

Отношение ассоциации является неиерархическим и ассоциативным. Ассоциативное отношение наиболее трудно определить. Российский стандарт на создание ИПТ указывает, что "ассоциативное отношение является объединением отношений, не входящих в *иерархические отношения* или в отношения синонимии. Допускается включать в ассоциативное отношение все виды отношений, кроме синонимии и отношения РОД-ВИД ".

Другие источники стараются изложить более подробные принципы установления ассоциативных отношений, поскольку в противном случае отношение будет проставляться непоследовательно.

Американский стандарт описывает наиболее общее правило установления ассоциативного отношения между дескрипторами таким образом: это отношение стоит устанавливать между двумя дескрипторами, если при употреблении одного термина другой термин как бы подразумевается. Один термин может быть необходимым элементом определения другого термина, например, термин клетка составляет необходимую часть определения термина цитология.

**Автоматическое индексирование по традиционным информационно-поисковым тезаурусам**

Поскольку основными элементами ИПТ являются термины, описанные как дескрипторы и аскрипторы, может показаться, что достаточно просто осуществить автоматическое индексирование *по* ИПТ путем простого сопоставления дескрипторов и аскрипторов с документами.

Однако для большинства документов такое автоматическое сопоставление не сможет отразить основное содержание документа:

* важные термины документа могут быть не найдены в тезаурусе, поскольку выражены в нем несколько иначе;
* менее значимые термины найдут прямое отражение в тезаурусе и выйдут на первый план и т.п.

Поэтому исследуются более сложные методы автоматизации индексирования *по* ИПТ.

Одним из подходов для автоматизации индексирования *по* традиционным ИПТ является подход, основанный на правилах. Такой подход к автоматическому индексированию был реализован *по* тезаурусу EUROVOC.

Правила могут быть простыми и сложными. Простые правила не содержат условий. Сложные правила содержат такие условия, как Близость (на расстоянии трех слов *по* тексту, в одном предложении, в том же самом *поле*, например, в *поле* реферата), Местонахождение (в заголовке, в тексте реферата или документа, в начале предложения, в конце предложения), Формат (с большой буквы, все большими буквами). Всего было создано около 40 тысяч правил.

В качестве других подходов автоматизации индексирования используются *статистические методы*.

При таких подходах процесс автоматического приписывания дескрипторов тезауруса EUROVOC полнотекстовым документам включает две стадии.

1. На первой стадии (этап обучения) на основе документов, вручную проиндексированных индексаторами, устанавливается соответствие между словами, встретившимися в тексте документа, и приписанными дескрипторами тезауруса. Соответствие устанавливается на основе статистических мер (chi-square или log-likelihood). Вес соответствия отдельного слова ключевому слову тем выше, чем выше совместная частотность использования данного слова и данного ключевого слова относительно частотности во всей коллекции.

Например, дескриптору тезауруса FISHERY MANAGEMENT соответствуют следующие слова (в порядке убывания веса): fishery, fish, stock, fishing, conservation, management, vessel и т.д.

1. На второй стадии (собственно индексирование) для каждого слова документа проверяется, каким дескрипторам тезауруса оно соответствует. Если такие дескрипторы имеются, то слово добавляет к весу дескриптора для данного текста натуральный логарифм веса, полученного на первом этапе. После обработки всех слов текущего текста получается суммированный вес дескрипторов тезауруса.

Например, для Резолюции по правам языковых и культурных меньшинств в Европейском союзе были получены следующие дескрипторы (в порядке убывания веса): Community programme, Young person, Cultural policy, CEEC, European Union и т.д.

Индексаторы Европейского Парламента присваивают документу обычно от 3 до 10 дескрипторов.

Выдачу системы можно ограничить *по* количеству выдаваемых дескрипторов или *по* весу. Для текста примера присвоенные индексаторами дескрипторы находились в первой тридцатке дескрипторов, присвоенных автоматически (на позициях 3, 8, 9, 16 и 30).

При этом большинство автоматически присвоенных дескрипторов выглядят весьма релевантными тексту документа и только 3 из 40 присвоенных автоматически явно неправильны (например, Кипр).

**Сочетание свободных запросов и запросов на основе информационно-поисковых тезаурусов**

В настоящее время в мире существует достаточно много информационных систем, предоставляющих пользователям возможности поиска информации как *по* свободному запросу на естественном языке, так и с помощью дескрипторов *информационно-поисковых тезаурусов*, сопоставленных документам профессиональными индексаторами.

Первым шагом на этом пути может быть нахождение корреляций между словами документов и дескрипторами тезауруса или рубриками рубрикатора.

Эксперименты *по* автоматическому расширению свободного запроса пользователя дескрипторами тезауруса проводились на двуязычной коллекции немецких и английских документов *по* общественным наукам. База включает в себя более 150 тысяч немецких документов и 26 тысяч английских. Документы реферативного характера содержат заголовок публикации, реферат и дескрипторы Тезауруса *по*общественным наукам, приписанных индексаторами. Эксперименты выполнялись в рамках предметно-ориентированного задания форума *по*многоязыковым информационным системам CLEF (Cross *Language* *Evaluation* *Forum*).