Автоматическая обработка больших массивов англоязычных текстов

Добров Б.В., Лукашевич Н.В.
Центр Информационных Исследований
{dobroff, louk}@mail.cir.ru

Описывается процесс автоматической обработки больших потоков англоязычных текстов широкой тематики. Ядром процесса является автоматическое индексирование англоязычных текстов дескрипторами двуязычного Тезауруса по общественно-политической жизни, автоматическое построение тематического представления англоязычного текста, описывающего темы и подтемы текста. На основе полученного тематического представления возможно проведение автоматической рубрикации текстов как по англоязычным, так и по русскоязычным рубрикаторам, построение аннотации, а также структурное изложение содержания текста на русском языке.

Введение

В связи с развитием Интернет русскоязычному пользователю становятся доступными огромные ресурсы англоязычных текстов, что делает необходимым развитие программных средств, помогающих широкому кругу русскоязычных пользователей найти нужную информацию, содержащуюся в англоязычных документах.

Такими средствами могли бы стать двуязычные информационные системы [1], способные по запросу на русском языке найти релевантные тексты на английском языке. К сожалению, существующие системы машинного перевода, способные порождать достаточно полный подстрочник, являются эффективными либо для весьма ограниченных предметных областей, либо только в режиме интерактивного взаимодействия с человеком, который выделяет правильные варианты из большого количества предлагаемых. Для больших потоков информации (десятки, сотни тысяч документов) эти обстоятельства не

позволяют добиться требуемой пользователями функциональности только с помощью систем полного машинного перевода.

Способом, позволяющим поднять общую эффективность многоязычных информационных систем, может являться использование программных систем, ориентированных на понимание основного содержания анализируемых текстов без полного перевода. Системы двуязычной рубрикации текстов, способные распределить англоязычные тексты по рубрикам, сформулированным на русском языке; системы двуязычного структурного аннотирования, передающие содержание англоязычного текста в структурной форме на русском языке, позволили бы русскоязычному пользователю отбирать нужные ему тексты для подробного чтения, машинного или ручного перевода.

Данная статья посвящена обработке англоязычных текстов, в процессе которой производится автоматическое индексирование англоязычных текстов дескрипторами двуязычного Тезауруса по общественно-политической жизни, строится тематическое представление англоязычного текста, описывающее темы и подтемы текста. На основе полученного тематического представления возможно проведение автоматической рубрикации текстов как по англоязычным, так и по русскоязычным рубрикаторам, построение аннотации, а также структурное изложение содержания текста на русском языке.

Построение тематического представления русскоязычных текстов общественно-политической области (официальных документов Российской Федерации, сообщений информационных агентств) и его использование для информационного поиска документов, автоматического рубрицирования и аннотирования подробно описано в работах [2 - 6].

В этой статье мы опишем развитие англоязычной части Тезауруса, которое было необходимо провести для эффективной обработки англоязычных текстов, видоизменения в алгоритме тематического анализа, которые были выполнены в связи с переходом на обработку англоязычных текстов.

Проведенная работа позволила нам принять участие в конференции TREC-6 [7] (см.п.8), в рамках которой было выполнено задание по поиску по 50 заданным темам на массиве англоязычных текстов объемом около 500 Мб (около 100 тысяч отдельных текстов, для каждой темы требовалось выдать упорядоченный по релевантности список 1000 наиболее релевантных документов).

1. Тематическое представление текста

Построение тематического представления текста базируется на свойствах локальной и глобальной связности текста и на предположении о том, что основная тема текста может быть описана некоторой пропозицией. В работе [8] такая пропозиция называется макропропозицией. Соответственно, назовем понятия, используемые в макропропозиции текста, макропонятиями, а слова и термины, используемые в соответствующей словесной формулировке основной темы текста макротерминами. Например, в тексте, основной темой которого является финансирование жилья военнослужащих, такими макропонятиями являются *ФИНАНСИРОВАНИЕ, ЖИЛЬЕ, ВОЕННОСЛУЖАЩИЙ.*

В работах [8, 9] указывается, что глобальная связность текста проявляется, прежде всего, в том, что основное содержание текста может быть представлено как иерархическая структура в том смысле, что тема всего текста может быть обычно описана посредством более конкретных тем текста, которые в свою очередь могут быть охарактеризованы посредством еще более конкретных подтем и т.п. Каждое предложение связного текста посвящено раскрытию той или иной подтемы основной темы текста.

Следствием предположения о глобальной связности текста является то, что, как правило, повторы макротерминов в тексте, использование семантически и тематически близких к макропонятиям слов и терминов, имеют непосредственное отношение к этим макропонятиям. Повторы макротерминов, ниже- и вышестоящие термины могут находиться в отношении кореференции или концептуального тождества с макропонятиями, тематически близкие понятия раскрывают аспекты макропонятия, обсуждаемые в тексте. Тематически близкими понятиями для макропонятия *ЖИЛЬЕ* в тексте могут быть, например, следующие: *ЖИЛИЩНОЕ СТРОИТЕЛЬСТВО, АРЕНДА ЖИЛЬЯ, КВАРТИРА, ОБЩЕЖИТИЕ*.

Таким образом, глобальная связность текста реализуется, в частности, посредством совокупностей терминов, семантически и тематически близких к макропонятиям. Совокупность понятий текста тематически близких одному и тому же понятию назовем тематическим узлом, а само это понятие тематическим центром.

Поскольку макропонятия в совокупности характеризуют основную тему текста, то можно считать, что глобально связанный текст посвящен описанию отношений между этими макропонятиями. Поэтому основным содержанием большинства подтем текста является описание отношений между элементами различных основных тематических узлов (тематических узлов вокруг макропонятий). Это значит, что пары терминов,

принадлежащих различным основным тематическим узлам, должны встречаться в тексте рядом чаще, чем термины тематических узлов, построенных вокруг других понятий текста.

Отсюда следует, что элементы тематических узлов каждого макропонятия должны проходить “тематическими нитями“ через весь текст и постоянно упоминаться в различных сочетаниях рядом друг с другом. Это постоянное совместное упоминание и есть та особенность тематических узлов макропонятий, которая выделяет их среди других возможных тематических узлов текста и позволяет находить их в текстах автоматически с большой точностью.

Эта закономерность определяется свойствами связности текста и не зависит от языка, на котором написан текст.

В качестве примера рассмотрим фрагменты англоязычного текста из текстового массива конференции TREC-6 (FB6-F001-0015). Текст описывает особенности борьбы российских пограничных войск с незаконной ловлей рыбы японскими рыбаками в российских территориальных водах.

Обозначим макропонятия текста и тематически связанные с ними термины следующим образом:

R	Россия	Russia, Russian, President of Russia, Southern Kurile, country
T	пограничные войска	border troops, border guards
P	браконьерство	poaching, poacher, poach, illegal activities
F	рыба	fish, fisherman, fishing, catch fish
J	Япония	Japan, Japanese, country
W	территориальные воды	territorial waters
B	судно	boat, schooner

используем обозначение **{ / }** для терминов , тематически близких к нескольким макропонятиям**,** ‘x‘ для других терминов Тезауруса, найденных в приводимых фрагментах текста, ‘_’ - для нетерминов.

Запишем фрагменты текста, используя введенные обозначения (Рис.1).

Элементы тематических узлов каждого макроаргумента проходят “тематическими нитями “ через весь текст и постоянно упоминаются в различных сочетаниях рядом друг с другом. Полученные выводы могут быть отражены в так называемом тематическом представлении текста.

“Border Troops `Putina' *Exercise* to *Control* **Poaching**	**Tt** _ x x **P**.
The **border troops** "are not saber rattling" in **Russian territorial waters** in the *Far East* as the *mass media*, especially the **Japanese** *mass media*, are attempting to portray it.	**Tt** _ _ **R Ww** Xx Xx, _ **J** Xx, _ _ _.
Servicemen have been legally granted the right to utilize all of the tools at their disposal, including *weapons*, to put a stop to **poaching**.	x _ _ _ _ _ _ _ _ _, _ x, _ _ **P**.
Russian Border Troops *Commander-in-Chief Colonel-General* Andrey Nikolayev stated that to an ITAR-TASS *correspondent* while stressing that his subordinates are conducting a strict policy to put a stop to the **illegal activities** of foreign **boats**.	**R Tt** Xx Xx _ _ _ _ _ x _ _ _ _ _ _ _ _ **Pp** _ **B**.
He noted that the **President of Russia** supports the position of the **border troops** for the full observance of the *law* in the **country's territorial waters**.”	_ _ **Rr** _ _ **Tt** _ _ x **{R/J} Ww**.
....	
“Incidentally, **Japanese fishermen** (read **poachers**) have learned about the operation beforehand that is called upon to put pressure on them.	_, **J F** (_ **P**) _ _ _ _ _ _. _ _ _ _ _ _ _ _.
This is certainly how we can explain why they have recently stepped up their activities.	_ _ _ _ _ _ _ _ _ _ _ _.
So, just from 26 March through 1 April and only in the **Southern Kurile** direction (Izmena Strait and Tanfilyev and Anuchin *islands*), 49 **Japanese boats** undertook attempts to **poach**.	_ _ _ _ _ **Rr** _ (_ _ _ _ x), _ **J B** _ _ **P**.
The **schooners** penetrated up to 55 cable lengths (one cable length is approximately 200 meters) into **Russian territorial waters**.”	**B** _ _ _ (_ _ _ _ _) **R Ww**.
....	
“But then again, the problem is much broader than just putting a stop to **poaching** in our **territorial waters**.	_ _ _ _ **P** _ **W w.**
It is also whether or not we, having established *monitoring* of **fishing**, will be able to conduct our own **fishing** for **fish** and *crabs* in these waters using our own men and *equipment*, in these waters that have been designated by and that are so familiar for **Japanese fishermen**?	_ _, _ x **F**, _ _ _ **F F** x _ _ _ _ _ x, _ _ _ **J F**?
That is not an idle question.	_ _ _.
The proposals to *sell* **fish** to the **Japanese** that are being increasingly loudly stated today are certainly well thought out. How?	_ x **F J** _ _ _ _ _ _. _?
Quite legally:	**_ _:**
By increasing their quota to **catch fish** in **Russian territorial waters** for *hard currency*.	**_ _ _ Ff R Ww Xx.**
In a word, even in this case the **border guards** will not be standing idly by:	_, _ **Tt** _ _.
Along with **fish** conservation *personnel*, they could carry out *monitoring* and continue to defend **Russia's** *economic* interests in the region.”	_ **F** _ x, _ _ x _ _ **R** x _ _.

Рис.1. Фрагменты текста TREC-6 (FB6-F001-0015)

Тематическое представление текста - это иерархическая структура терминов текста, в которой тематически близкие термины собраны вокруг тематических центров в тематические узлы, среди тематических узлов выделены:

Тематические узлы связаны между собой отношением *иметь_отношение_к*. Тематические узлы можно классифицировать по суммарной частотности терминов их составляющих, а также по суммарной текстовой связности с другими узлами:

- основные тематические узлы, отражающие в совокупности основное содержание всего текста;

- локальные тематические узлы, отражающие подтемы текста.

Иерархия тематического представления отражает важность для текста тех или иных терминов. Тематический центр значимее других терминов тематического узла, термины основных тематических узлов более значимы для текста, чем термины других тематических узлов.

Основой для построения тематического представления текста служит Тезаурус по общественно-политической жизни (далее Тезаурус) [2], специально разработанный как средство, используемое в автоматических процессах обработки текстов.

3. Тезаурус для автоматического индексирования

Тезаурус для автоматического индексирования, как и тезаурус для ручного индексирования [12, 13, 14], остается по своему назначению контролируемым словарем индексирования. Однако принципы построения тезауруса для автоматического индексирования существенно отличаются от принципов построения тезауруса для ручного индексирования.

Тезаурус для ручного индексирования представляет собой язык-посредник для описания основного содержания текстов предметной области. Во всех известных авторам статьи тезаурусах для ручного индексирования неявно учитывается, что объем языковых знаний, в том числе лексических и терминологических, на основе которых специалист-индексатор принимает решение об основном содержании того или иного документа, значительно шире. Поскольку при автоматическом индексировании место человека-индексатора занимает автоматический процесс, то тезаурус для автоматического индексирования должен включать в себя значительно более подробное и точное описание лексических и терминологических знаний.

Прежде всего, для улучшения распознавания понятий Тезауруса в тексте значительно увеличивается описываемая совокупность вариантов понятия, многие из которых могут показаться избыточными (и действительно такими являются) для человека-индексатора. Например,

КАССАЦИОННОЕ ПРОИЗВОДСТВО

кассационное опротестование
кассационный порядок
кассационный протест
кассация приговора
кассация судебного решения
обжалование в кассационном порядке
производство в кассационной инстанции

Тезаурус для автоматического индексирования должен включать средства описания многозначных терминов, позволяющие в процессе автоматической обработки текстов разрешать многозначность терминов.

Увеличивается степень подробности включенных понятий:

- практически без ограничений в Тезаурус для автоматического индексирования включаются конкретные понятия, идентификация которых в тексте может улучшить точность автоматического распознавания основной темы текста;

- возможно описание в Тезаурусе близких по смыслу понятий, что обычно не делается в тезаурусах для ручного индексирования в целях уменьшения фактора субъективности индексирования.

Существенным образом меняются цели описания связей между различными понятиями в Тезаурусе для автоматического индексирования. Описания связей между понятиями необходимы:

- для автоматического расширения запроса пользователя в целях обеспечения полноты поиска;

- для нахождения тематически близких терминов в тексте, что необходимо для идентификации основной темы текста и выбора понятий, наиболее точно отражающих выявленную основную тему;

- для разрешения многозначности терминов.

Изменение целей описания отношений, состава описываемых понятий и использование в автоматическом режиме повлекло изменение номенклатуры тезаурусных

отношений. Обычный тезаурусный набор отношений между понятиями (ВЫШЕ-НИЖЕ и АССОЦИАЦИЯ) был расширен, однако не в сторону увеличения содержательности отношений, а в сторону описания возможностей автоматического вывода для решения поставленных целей.

Одновременно Тезаурус существенно отличается и от общеязыковых тезаурусов типа WordNet [10, 11]. Часть этих различий связана с тем, что Тезаурус специально создавался как инструмент для автоматической обработки текстов, часть различий имеет свои корни в том, что Тезаурус представляет собой описание понятийной модели конкретной, хотя и очень широкой предметной области

Эти различия прежде всего проявляются в формировании словарного состава Тезауруса. Поскольку Тезаурус является тезаурусом в общественно-политической области, а не описывает понятийную структуру языка в целом, то

- в Тезаурус не включаются наиболее общезначимые, (и часто это наиболее многозначные слова), которые могут функционировать в любой подобласти языка;

- в Тезаурус не включаются те значения терминов, которые очень редко встречаются в текстах предметной области и эти термины могут считаться однозначными;

- в Тезаурусе не описываются те значения многозначных терминов, которые отражают общезначимые понятия, вместо этого ставится пометка о многозначности, что означает, что нужно провести дополнительные проверки для подтверждения того, что термин в анализируемом тексте употреблен именно в описываемом в Тезаурусе значении.

Существенно различаются принципы включения в Тезаурус многословных терминов. Для общеязыкового тезауруса процесс включения в состав многословного языкового выражения во многом связан с описанием отдельных слов (словосочетания являются синонимами отдельных слов, служат для поддержания понятийной иерархии значений отдельных слов и т.п.).

Принципы включения в Тезаурус многословных терминов тесно связаны с его использованием в автоматических процедурах. Важнейшими принципами включения в Тезаурус многословных терминов являются следующие:

- существование у словосочетания связей с другими элементами Тезауруса, не выводимых из связей каждого отдельного слова, входящего в состав этого слововсочетания (*землепользование - аренда земли*);

- вхождение в состав словосочетания многозначного слова, которое в словосочетании имеет ровно одно значение;

- совокупность словосочетаний некоторого термина Тезауруса с общезначимыми словами представляет собой практически варианты одного и того же понятия *(хирургическая операция - хирургическое вмешательство*).

3.1 Развитие Тезауруса по общественно-политической жизни для обработки англоязычных текстов

Процесс развития Тезауруса по общественно-политической жизни из русскоязычного в двуязычный включает в себя несколько этапов.

Значительная часть из существующих более 18500 дескрипторов Тезауруса (6500 из них - имена собственные, в основном, географические названия) и частично их варианты были переведены на английский язык. Среди множества полученных переводов, связанных с тем или иным дескриптором, был выбран один, который стал названием английского дескриптора. В качестве вариантов дескриптора включались различные варианты (например, английский и американский) написания самого дескриптора и его синонимов.

Как и в русскоязычной части Тезауруса, английские термины могли быть вариантами для различных дескрипторов, однако сами дескрипторы должны были быть уникальными, что достигалось либо поиском синонимичных однозначных терминов, или включением в состав названия дескриптора дополнительных помет. Система связей между дескрипторами русскоязычного тезауруса была сохранена и в англоязычной части Тезауруса.

Были дополнительно изучены англоязычные Тезаурусы по сходной тематике [12,14]. С их помощью либо дополнялся список вариантов англоязычного дескриптора, либо менялось само название англоязычного дескриптора, как наиболее общепринятое, лучше отражающее соответствующее понятие.

Однако чтобы сделать Тезаурус средством, пригодным для использования в автоматической обработке англоязычных текстов, необходимо было выполнить еще несколько процедур.

Прежде всего оказалось, что в результате описанной процедуры перевода, количество англоязычных терминов, не являющихся собственными именами, вдвое меньше соответствующего количества русскоязычных терминов (Рис.2).

РЫБОЛОВСТВО		**fishing**	
	UF ВЫЛОВ РЫБЫ; ДОБЫЧА РЫБНЫХ РЕСУРСОВ; УЛОВ РЫБЫ; ДОБЫЧА РЫБЫ; ЛОВ РЫБЫ; ПРОМЫСЕЛ РЫБЫ; ПРОМЫСЛОВОЕ РЫБОЛОВСТВО; ПРОМЫСЛОВЫЙ ЛОВ; ПРОМЫШЛЕННОЕ РЫБОЛОВСТВО; РЫБНАЯ ЛОВЛЯ; РЫБНЫЙ ПРОМЫСЕЛ; РЫБОДОБЫВАЮЩИЙ; РЫБОЛОВНЫЙ; РЫБОЛОВЕЦКИЙ; РЫБОЛОВНАЯ ДЕЯТЕЛЬНОСТЬ; РЫБОПРОМЫСЛОВЫЙ; РЫБОПРОМЫСЛОВАЯ ДЕЯТЕЛЬНОСТЬ	**UF**	**catch fish**
BT	ВОДНЫЙ ПРОМЫСЕЛ **UF** ПРОМЫСЕЛ ВОДНЫХ БИОРЕСУРСОВ	**BT**	**fishery**
NT	МОРСКОЕ РЫБОЛОВСТВО **UF** ОКЕАНИЧЕСКОЕ РЫБОЛОВСТВО	**NT**	**maritime fishery**
NT	НЕЗАКОННЫЙ ЛОВ РЫБЫ	**NT**	**illegal fishing**
NT	ПРЕСНОВОДНОЕ РЫБОЛОВСТВО **UF** ПРУДОВОЕ РЫБОЛОВСТВО	**NT**	**freshwater fishing**
NT	ТРАЛОВЫЙ ЛОВ **UF** ТРАЛОВАЯ ОПЕРАЦИЯ; ТРАЛОВЫЙ ПРОМЫСЕЛ	**NT**	**trawl fishing** **UF trawling**
NT	ЛЮБИТЕЛЬСКОЕ РЫБОЛОВСТВО **UF** ЛЮБИТЕЛЬСКАЯ ЛОВЛЯ; ЛЮБИТЕЛЬСКИЙ ЛОВ		**--**
PT	РЫБАК **UF** РЫБОЛОВ	**PT**	**fisherman**
PT	РЫБОЛОВНОЕ ПРЕДПРИЯТИЕ **UF** РЫБКОЛХОЗ; РЫБОДОБЫВАЮЩАЯ ОРГАНИЗАЦИЯ; РЫБОДОБЫВАЮЩЕЕ ПРЕДПРИЯТИЕ; РЫБОЛОВЕЦКАЯ АРТЕЛЬ; РЫБОДОБЫВАЮЩИЙ ТОВАРОПРОИЗВОДИТЕЛЬ; РЫБОЛОВЕЦКИЙ КОЛХОЗ; РЫБОЛОВЕЦКОЕ ПРЕДПРИЯТИЕ; РЫБОЛОВНАЯ ОРГАНИЗАЦИЯ; РЫБОЛОВНОЕ ХОЗЯЙСТВО; РЫБОПРОМЫСЛОВАЯ ОРГАНИЗАЦИЯ	**PT**	**commercial fishery enterprise**
PT	РЫБОЛОВНЫЕ ОРУДИЯ **UF** ОРУДИЕ ЛОВА; РЫБОЛОВНАЯ СНАСТЬ; РЫБОЛОВНОЕ СНАРЯЖЕНИЕ	**PT**	**fishing equipment**
PT	РЫБОПРОМЫСЛОВАЯ РАЗВЕДКА	**PT**	**fish reconnaisance**
PT	РЫБОПРОМЫСЛОВЫЙ ФЛОТ **UF** ПРОМЫСЛОВЫЙ ФЛОТ; РЫБНЫЙ ФЛОТ; РЫБФЛОТ; РЫБОЛОВЕЦКИЙ ФЛОТ; РЫБОЛОВНЫЙ ФЛОТ; ТРАЛОВЫЙ ФЛОТ; ФЛОТ РЫБНОЙ ПРОМЫШЛЕННОСТИ	**PT**	**fishing fleet**
RT	РЫБА **UF** ВИД РЫБ; РЫБНОЕ СЫРЬЕ; РЫБНЫЙ	**RT**	**fish**
RT	РЫБНАЯ ПРОДУКЦИЯ **UF** МЯСО РЫБЫ; РЫБНАЯ ГАСТРОНОМИЯ; РЫБОТОВАРЫ РЫБНЫЕ ПРОДУКТЫ; РЫБНЫЕ ТОВАРЫ; РЫБОПРОДУКТЫ; РЫБОПРОДУКЦИЯ	**RT**	**fish products**
RT	РЫБНЫЕ РЕСУРСЫ **UF** РЫБНЫЕ ЗАПАСЫ	**RT**	**fish resources** **fish wealth**

Рис 2. Пример тезаурусной статьи

Анализ текстов дал возможность дополнить ряды англоязычных вариантов для дескрипторов, обнаружить варианты написания терминов. В качестве примера приведем (Рис.3) следующую таблицу (первый столбец - русское название дескриптора, второй столбец - английское название дескриптора, третий столбец - вариант перевода, обнаруженный в текстах).

Русский дескриптор	**Перевод в Тезаурусе**	**Новый вариант**
нефтяник	oil fielder	oil worker
строительство домов	house construction	construction of buildings
сейсмическая безопасность	seismic security	seismic safety
сточные воды	sewage	wastewater
земельный участок	land lot	parcel
экологическая политика	ecological policy	environmental policy
министр обороны	minister of defence	defence minister
деревообрабатывающая промышленность	wood processing industry	wood products industry
водоплавающие птицы	water-fowl	waterfowl
перестрахование	re-insurance	reinsurance

Рис.3.

Реальной проблемой для применения англоязычного тезауруса для анализа английских текстов стало то, что многие однозначные или практически однозначные в предметной области русскоязычные термины переводятся на английский язык общезначимыми и очень многозначными словами, причем для английского языка характерна очень обширно представленная неоднозначность определения части речи многих слов (Рис.4).

Русский термин	**Один из вариантов перевода**
инвалид	invalid
концерн	concern
термин	term
свинец	load
майор	major
трюм	hold
завещание	will
родственник	relative
судебное дело	case

Рис.4.

Обилие такого рода переводов привело к необходимости ужесточить процедуру разрешения морфологической неоднозначности терминов (см. п. 4.2).

Для повышения эффективности разрешения многозначности английских терминов необходимо выявлять по текстам такие варианты дескрипторов Тезауруса, в которых многозначные термины становятся однозначными. Эта работа активно проводилась и для русскоязычной части Тезауруса.

Приведем примеры таких словосочетаний:

land use

agricultural *land*

preservation of *land*

lease of *land*

credit mechanizm

environmental contamination

granting benefits

Работа по расширению англоязычной части Тезауруса еще продолжается.

4. Построение тематического представления для англоязычных текстов

4.1 Нахождение терминов в тексте

На первом этапе работы алгоритма единицы текста сравниваются с единицами Тезауруса. Это сравнение происходит на основе морфологического представления единиц текста и единиц Тезауруса. Для этого был разработан упрощенный морфологический анализатор англоязычных текстов.

Морфологический анализатор основан на применении 11 основных правил отсечения окончаний английских слов и 12 тысячах исключений, источником которых явился словарь WordNet [10]. В процессе своей работы анализатор использует словарь из 500 наиболее частотных слов, заданных заранее, и буфер на 20000 словоформ, в котором содержатся результаты предшествующих морфологический разборов слов.

При переполнении буфера наиболее редко встречающиеся словоформы удаляются, а частотность остальных понижается с некоторым коэффициентом, что позволяет осуществлять поднастройку буфера на новый массивы текстов. Результаты морфологического анализа, хранящиеся в буфере, могут быть просмотрены и подправлены вручную, подправленные результаты разбора далее не удаляются.

Очередная словоформа последовательно ищется в списке наиболее частотных слов, буфере и в словаре исключений, и лишь если она нигде не найдена, производится ее разбор.

Все термины Тезауруса предварительно получают свое морфологическое представление в виде совокупности основных форм, входящих в их состав терминов. Для каждого поступающего текста проводится морфологический анализ. Если один и тот же фрагмент текста соответствует разным единицам Тезауруса, то фиксируется многозначность термина.

В результате сопоставления с Тезаурусом текст отображается в последовательность дескрипторов Тезауруса. Все синонимы (варианты) одного и того же дескриптора отображаются в соответствующий дескриптор и далее не различаются. Для каждого дескриптора фиксируется частота его встречаемости в тексте и тематически близкие ему дескрипторы текста.

Совокупность дескрипторов текста, для которых указаны тематически близкие дескрипторы этого текста, называется проекцией Тезауруса на текст (тезаурусной проекцией).

Параллельно производится построение текстовых связей для каждого дескриптора текста, т.е. фиксация для каждого вхождения каждого дескриптора трех соседних дескрипторов вправо и трех влево. Выбор именно таких параметров подтвержден достаточно значительными нашими экспериментами по обработке текстов, однако согласуется и с экспериментами в области исследования кратковременной памяти**.**

4.2 Разрешение неоднозначности терминов

В построении тезаурусной проекции равным образом участвуют все дескрипторы, соответствующие неоднозначному термину. На основе тезаурусной проекции выбирается дескриптор, соответствующий определенному значению термина. Для каждого значения неоднозначного термина проверяется:

- употреблялись ли в данном тексте наряду с неоднозначным термином однозначные термины, соответствующие дескриптору, выражающему это значение неоднозначного термина;
- имеет ли дескриптор, соответствующий этому значению неоднозначного термина, тезаурусные связи с другими дескрипторами проекции.

Если выполняется одно из перечисленных выше условий, то считается, что "текст поддерживает" данное значение неоднозначного термина. Для русскоязычного текста предполагалось, что если текст "поддерживает" только одно значение неоднозначного термина, то выбирается соответствующий ему дескриптор. Точность процедуры разрешения

неоднозначности терминов для русскоязычного текста превышала 75 процентов. Процедура работает особенно эффективно, если разные значения термина соответствуют различным подобластям внутри сферы общественно-политической жизни.

Прямое применение такой же процедуры к англоязычному тексту оказалось значительно менее эффективным из-за большого количества общезначимых слов совпадающих с англоязычными терминами Тезауруса. Например, в тексте примера употребляется термин *fine* в значении *штраф*, в месте с тем, лексема *fine* имеет также различные общезначимые значения как прилагательное. Поэтому для разрешения неоднозначности англоязычного термина было введено более жесткое правило: даже если текстом поддержано только одно значение термина, для выбора этого значения необходимо, чтобы поддерживающие термины находились на расстоянии текстовой связи от многозначного термина.

Если текст "поддерживает" дескрипторы, соответствующие разным значениям термина, то при обработке как русскоязычных, так и англоязычных текстов, для каждого вхождения неоднозначного термина рассматриваются ближайшие по тексту дескрипторы, для них проверяются указанные условия и выбирается тот дескриптор неоднозначного термина, который "поддерживается" первым из ближайших по тексту дескрипторов.

4.3 Построение тематических узлов и определение их статуса

Чтобы найти в тексте основные тематические узлы, сначала необходимо построить совокупность различных тематических узлов, каждый из которых может соответствовать как основным темам текста, так и подтемам. Мы предполагаем, что тот термин, который наиболее точно характеризует развиваемую в тексте тему и который соответственно может стать тематическим центром одного из тематических узлов текста, обычно некоторым образом выделяется в пространстве всех тематически близких терминов, а именно: такой термин может быть употреблен в заголовке и/или в начале текста, иметь максимальную частотность среди других тематически близких терминов. Тематическим центром может стать любой термин Тезауруса, независимо от уровня его общности/специфичности.

В процессе обработки текста создание тематического узла начинается с выбора главного дескриптора тематического узла. Сначала тематические узлы собираются вокруг дескрипторов заголовка и первого предложения текста. Затем тематические узлы собираются для остальных дескрипторов, начиная с самых частотных. Те дескрипторы,

которые уже попали в тематический узел некоторого дескриптора, свой тематический узел не образуют.

Приведем примеры тематических узлов, созданных в процессе обработки текста, приведенного в качестве примера в п. 2 (главный дескриптор тематического узла выделен сдвигом влево; указана также частота употребления дескриптора в тексте):

Russia (Russian)	10
Far East	1
Curile	1
President of Russia	1
state (country)	6
territorial waters	9
ocean	3
water transport (ship, schooner)	11
island	1
state (country)	6
fish	10
fish resources	1
natural resources	3
fishing (catch fish)	5
fishing vessel	3
fisherman	2
illegal fishing	1
pouching (pouch)	5
offence (illegal activity, violence)	3
illegal fishing	1
fisherman	2

После того как созданы тематические узлы, текстовые связи дескрипторов каждого тематического узла суммируются и определяются текстовые связи между тематическими узлами. Приведем примеры текстовых связей между тематическими узлами, выделенными в тематическом представлении текста примера. Число слева - суммарная величина текстовых связей между дескрипторами тематических узлов, текстовые связи даны для тематического узла, который смещен влево:

fish, fisherman, fishing, catch fish

10	*border troops. border guards*
10	*Japan, Japanese, country*
9	*territorial water*
7	*Russian Federation*
7	*poaching, poacher, poach, illegal activities*

В соответствии с моделью предполагается, что основными тематическими узлами в первую очередь являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями;

- сумма частот текстовых связей между ними максимальна.

Таким образом, “тематические нити”, образованные основными тематическими узлами проходят, взаимно переплетаясь через весь текст.

В рассматриваемом примере тематического представления основными тематическими узлами стали узлы с главными дескрипторами *border guards, territorial waters, Russian Federation, fish, poaching, Japan, control*.

Мы различаем также локальные тематические узлы, которые представляют собой некоторые важные характеристики основных тематических узлов. Тематический узел считается локальным, если этот узел имеет текстовую связь с частотностью большей единицы с одним из основных тематических узлов. По построению следует, что “тематические нити” локальных тематических узлов переплетаются с “тематическими нитями” основных тематических узлов только в некоторых фрагментах текста.

В нашем примере локальные тематические узлы были например, такие:

logistics *mass media*

equipment *correspondent*

computer

Дескрипторы, не вошедшие в состав основных и локальных тематических узлов, объявляются "упоминавшимися" в тексте.

Упоминавшимися дескрипторами были *legislator, expert, ice situation ....*

Разбиение тематических узлов на основные и локальные задается разбиение дескрипторов на следующие пять классов по их важности для анализируемого текста:

- главные дескрипторы основных тематических узлов (основные темы);

- другие дескрипторы основных тематических узлов;

- главные дескрипторы локальных тематических узлов (локальные темы);

- другие дескрипторы локальных тематических узлов;

- упоминавшиеся дескрипторы.

Таким образом, построено тематическое представление текста, в котором все термины текста разбиты на тематические узлы. Тематические узлы подразделяются на основные, локальные и упоминавшиеся узлы. Между тематическими узлами фиксируются текстовые связи.

5. Автоматическое рубрицирование англоязычных текстов

В статьях [4, 15] описана система автоматического рубрицирования, работающая на базе Тезауруса и тематического представления текстов. Система способна рубрицировать тексты различных типов (официальные документы, сообщения информационных агентств, газетные статьи), ее легко можно настроить на новый рубрикатор и новые типы текстов, рубрицирование можно осуществлять сразу по нескольким рубрикаторам.

На основе описываемой технологии реализованы две системы автоматической рубрикации текстов, причем каждая имеет свои рубрикаторы. Одна рубрицирует российские официальные документы, используя два рубрикатора, одним из которых является рубрикатор исследовательской службы Библиотеки Конгресса США [12]. Вторая система рубрицирует сообщения нескольких информационных агентств: ИМА-пресс, ИТАР-ТАСС, WPS.

Рубрикаторы связываются с Тезаурусом посредством небольшого числа опорных терминов, рубрики остальных терминов выводятся по связям внутри Тезауруса, что стало возможным благодаря тщательной предварительной разработке тезаурусных связей, максимально полному отражению различных аспектов описываемых понятий.

Используя классы, присвоенные дескрипторам текста, можно проводить два вида рубрикации - безусловную и ранжированную. При безусловной рубрикации рубриками текста становятся только те рубрики, которые соответствуют главным дескрипторам основных тематических узлов. При ранжированной рубрикации каждому тексту ставятся в соответствие рубрики первых трех категорий с указанием веса этой рубрики.

Поскольку после построения тезаурусной проекции текста безразлично, на каком языке был написан исходный текст, то все эти рубрикаторы (или новые) могут быть использованы для рубрикации англоязычных текстов.

Так, в процессе автоматической ранжированной рубрикации текст примера получил следующие рубрики рубрикатора, используемого для рубрикации официальных документов Российской Федерации (в процентах указана оценка степени уверенности):

Водный транспорт	*90%*
Государственная граница и территориальное деление	*90%*
Оборона	*90%*
Охрана государственной границы	*90%*
Преступления	*90%*

Природные ресурсы	*90%*
Экология	*90%*
Рыбное хозяйство	*60%*
.....	
Зарубежные регионы	*30%*
Контроль и статистика	*30%*

По рубрикатору исследовательской службы Конгресса США этот же текст будет иметь такие рубрики:

Criminal justice	*90%*
Defense policy	*90%*
International affairs	*90%*
Marine resources	*90%*
Natural resources	*90%*
Transportation	*90%*
Water resources	*90%*
.....	

6. Автоматическое аннотирование англоязычных текстов

Построенное тематическое представление позволяет автоматически строить аннотации текстов [5]. Автоматическое построение аннотаций может быть произведено и для английских текстов [16].

Отметим основные принципы автоматического построения аннотаций с использованием тематического представления текста.

Получая тематическое представление текста, мы предположили, что текст посвящен описанию макропонятий, которые мы моделировали основными тематическими узлами. Поэтому наиболее информативными предложениями текста, по нашей модели, являются такие предложения, которые содержат по крайней мере два термина из двух различных основных тематических узлов. Наиболее полно может характеризует текст такая аннотация, которая состоит из предложений, содержащих пары терминов из всех основных тематических узлов.

В тексте может содержаться много предложений, содержащих пары представителей одних и тех же тематических узлов. Из таких предложений мы выбираем предложение, которое ближе к началу текста. Такой выбор позволяет улучшить связность получаемых аннотаций, поскольку именно в начале обсуждения очередного макропонятия основной темы, автор должен аккуратно связать его с предшествующим текстом. Так мы можем положиться на усилия автора текста написать связный текст.

Описание предметной области в Тезаурусе очень подробно, поэтому первое появление в тексте очередного макропонятия может быть сразу идентифицировано.

Приведем пример автоматически полученной аннотации для текста примера (Рис.5).

Border Troops ` Putina ' Exercise to Control Poaching border troops " are not saber rattling " in Russian territorial waters in Far East as mass media, especially Japanese mass media, are attempting to portray it... Recently, we have become accustomed to reports on entry of Japanese fishing boats into Russian territorial waters to poach fish... And although Russian border guards, who are experiencing great difficulties in logistics - technical support due to well - known economic situation in country , have been able to observe approximately 140 foreign fishing boats and to fine poachers sum of more than 21 million rubles and over 100, 000 U.S. dollars in 1990, so far, their efforts are drop in sea.

Рис.5.

7. Аннотирование англоязычного текста русскоязычными средствами

То, что аннотация англоязычного текста составлена на английском языке, может затруднять ее восприятие русскоязычным пользователем, недостаточно знакомым с английским языком. Подобные проблемы, когда трудно составить понятную и связную аннотацию из фрагментов исходного текста, возникают также для таких документов как законодательные акты, имеющие очень сложную структуру и громоздкие предложения, газетные интервью, тексты большой длины и некоторые другие. Для представления содержания таких текстов предлагается структурная тематическая аннотация текста [6, 17], содержащая наиболее информативные фрагменты тематического представления. Структурная аннотация может конструироваться как для русскоязычных, так и для англоязычных текстов и может быть показана как на английском, так и на русском языках.

Структурная тематическая аннотация включает в себя следующие части:

- основные и локальные тематические узлы, приводимые в виде списков составляющих их терминов, упорядоченных по убывающей частотности и расположенные горизонтально;
- отметки об относительно суммированной частотности основных тематических узлов, обозначаемые различным количеством символов “*”;

- отметки о силе взаимоотношений между различными тематическими узлами

- ”X“-- очень сильное отношение;

- ”z”-- сильное отношение;

- ”.” -- отношение.

Такая аннотация позволяет оценивать содержание текста с первого взгляда.

Для текста примера русскоязычная структурная аннотация выглядит следующим образом (Рис.7).

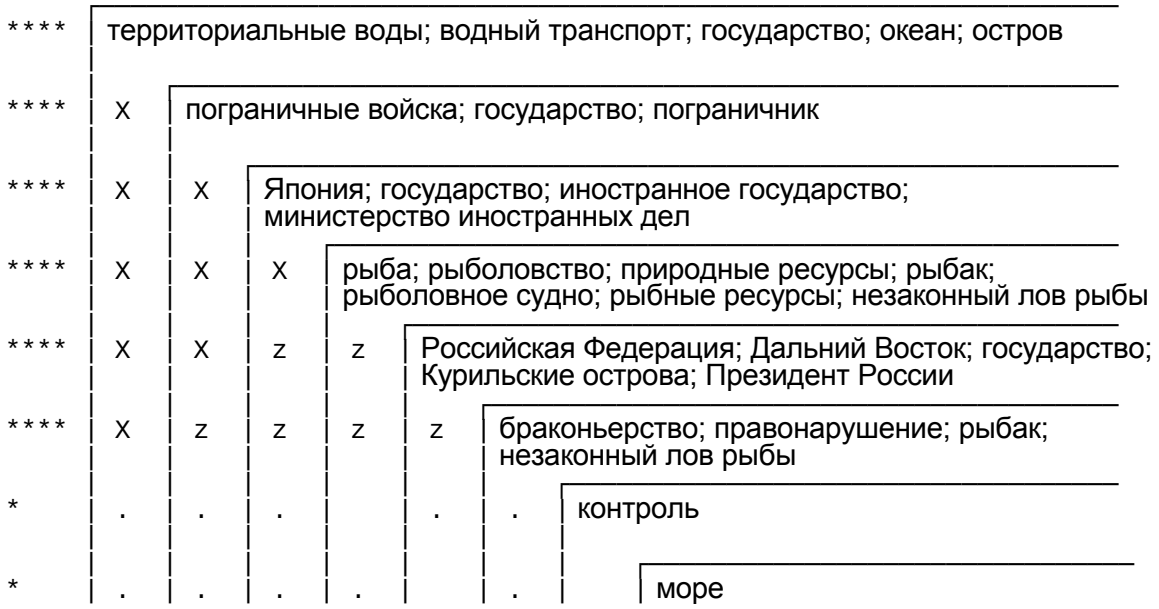

Рис.6.

8. Участие в конкурсе по обработке больших массивов англоязычных текстов

Технология автоматической рубрикации, опробованная на русскоязычных и англоязычных текстах, была применена для информационного поиска в большом массиве англоязычных текстов при участии в конференции TREC- 6 (Text Retrieval Conference) [17].

Конференция TREC проводится ежегодно с 1992 года. Основными целями ее проведения являются следующие:

- расширение исследований по информационному поиску в очень больших массивах документов;

- расширение коммуникации между научными, коммерческими и правительственными кругами в сфере информационного поиска;

- совершенствование процесса передачи технологии от исследования к конечному продукту;

- улучшение методов оценки эффективности информационного поиска.

Организаторы TREC'а подготовили громадный массив разнородных текстов (около 5 Гигабайт), состоящий из газетных статей (Wall Street Journal, Financial Times), сообщений информационных агентств (the Associated Press) и др. Размеры документов варьируются от 300-400 байт в длину до нескольких сотен страниц.

Документы рассылаются участвующим группам. исследователей Далее задается некоторое множество запросов, которые участвующие группы должны выполнить на заданном массиве документов (в настоящее время до 2 Гигабайт), определить 1000 документов, наиболее соответствующих каждому запросу, упорядочить найденные документы по релевантности теме, сообщить их организаторам конференции к назначенному сроку (обычно 2 месяца).

Организаторы собирают все найденные документы по каждому запросу и передают их экспертам для оценки релевантности каждого документа запросу. После завершения этого процесса каждой участвующей группе сообщаются оценки эффективности выполненного ею поиска и проводится трехдневная конференция, на которой участники представляют свои системы.

В 1996 на основе технологии автоматической рубрикации нами был выполнен поиск в англоязычном текстовом массиве величиной 500 Мб по 50 темам [6].

Каждая тема была описана как логическое выражение:

$$\bigsqcup X_i = \bigsqcup (\& x_{ij}) .$$

Каждому операнду x_{ij} были сопоставлены опорные дескрипторы из Тезауруса. После этого список дескрипторов, соответствующих каждому операнду, был расширен за счет нижестоящих дескрипторов.

В результате,

$$x_{ij} = \bigsqcup w_{ijk} \quad ,$$

где w_{ijk} дескрипторы из Тезауруса.

Например, запрос по теме 012 "Water Pollution - documents is about the pollution of a body of water." (Загрязнение вод) был определен как:

$$X_1 \bigsqcup X_2 = \text{A} \bigsqcup (\text{B} \& \text{C})$$

$$X_1 = x_{11}; \; x_{11} = \text{A}; \quad X_2 = (x_{21} \& x_{22}); \; x_{21} = \text{B}; \; x_{22} = \text{C}$$

Более подробное описание этой темы показано на рис.7.

	x_{ij}	w_{ijk}
012	**A** **“water pollution”**	federal water pollution control act; federal water pollution control administration; hot water pollution; sewage disposal pollution of sea environment; sewage water pollution; water purification water supply and pollution control division
012	**B** **“pollution”**	ground pollution; oil distribution supertanker shipwreck oil pollution; oil spill
012	**C** **“body of water”**	body of water; animalis aquaticus; basin; fresh water; freshwater fishing; freshwater aquaculture; inland waterways freshwater reservoir; maritime fishery; lake ocean; ocean resources; reservoir; river; salt water; sea; sea animal; sea fish sea flora; sea mammal; sea-water; water basin sources of water; surface waters; water biological resources; water plant water resources; water scoop; water supply water-way; watershed

Рис. 7. Описание темы TREC6-012

“Water Pollution - documents is about the pollution of a body of water.” (Загрязнение вод)

В процессе обработки каждого документа мы вычисляли вес соответствия этого документа каждой заданной теме.

Основное правило вычисления веса было следующим:

$$\mu_D = \max_i(\mu_X(X_i)) \ ,$$

где вес каждой группы операндов:

$$\mu_X(X_i) = \Pi_j \mu_x(x_{ij}) = \mu_x(x_{i1}) \cdot \mu_x(x_{i2}) \cdot .... \cdot \mu_x(x_{im}) \ ,$$

и вес операнда вычисляется как:

$$\mu_a(x_{ij}) = max\{\mu_0 , \nu_T(w_{ijk})\} , \qquad \mu_0 = 0.001 ,$$

$$\nu_T(a_{ij}) = \begin{cases} 1.00, & \text{if } a_{ij} \text{ - главный дескриптор основного тематического узла,} \\ 0.60, & \text{if } a_{ij} \text{ - другой дескриптор основного тематического узла,} \\ 0.30, & \text{if } a_{ij} \text{ - главный дескриптор локального тематического узла,} \\ 0.10, & \text{if } a_{ij} \text{ - другой дескриптор локального тематического узла,} \\ 0.05, & \text{if } a_{ij} \text{ - упоминавшийся дескриптор,} \\ 0.00, & \text{иначе.} \end{cases}$$

После построения тематического представления, когда имеется достаточно много информации о содержании документа, можно предложить и другие формулы оценки веса термина для содержания документа.

Заключение

Проведенные эксперименты по автоматической обработке англоязычных текстов показали, что процесс построения тематического представления текста после сопоставления текста с Тезаурусом не зависит от языка написания текста, и также как для русскоязычных текстов [2-6] тематическое представление англоязычных текстов может служить основой для их автоматического рубрицирования и аннотирования [15-17]. В будущем планируется проводить дальнейшую работу по поиску новых англоязычных вариантов дескрипторов Тезауруса, совершенствовать морфологический анализ англоязычных слов, процедуру разрешения неоднозначности англоязычных терминов.

СПИСОК ЛИТЕРАТУРЫ

1. Oard D., Dorr B. 1996. A Survey of Multilingual Text Retrieval. Technical report, UMIACS-TR-96-19 CS-TR-3615.
2. Лукашевич Н.В., Салий А.Д. Представление знаний в системе автоматической обработки текстов // НТИ. Сер.2. - 1997 - N3.
3. Лукашевич Н.В., Добров Б.В. Построение и использование тематического представления содержания документов // 5 Национальная конференция КИИ-96. - Казань, 1996. - С.130-134.
4. Лукашевич Н.В. Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. - 1996. - N 10. - С.22-30.
5. Лукашевич Н.В. Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. - Москва, 1997. - С.188-191.
6. Добров Б.В., Лукашевич Н.В. Построение структурной тематической аннотации текста // Труды международного семинара Диалог-98, Том 2 - Казань, 1998 - С.795-802.
7. Dobrov B., Loukachevitch N., Yudina T. Conceptual Indexing Using Thematic Representation of Texts // Information Technology: The Sixth Text Retrieval Conference (TREC-6), Ed. E.M.Voorhees, D.K.Harman - NIST Special Publication 500-240, 1998 - pp.403-413.
8. ван Дейк Т.А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. Вып. 23. - М.: Прогресс, 1988. - С.153-211.

9. Леонтьева Н.Н. О компонентах системы понимания текста // Уровни текста и методы его лингвистического анализа . М., 1982, С.124-140.
10. Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Five papers on WordNet. CSL Report 43.Cognitive Science Laboratory, Princeton University. - 1990.
11. Climent S., Rodriguez H., Gonzalo J. Definitions of the links and subsets for nouns of the EuroWordNet project. - Deliverable D005, WP3.1, EuruWordNet, LE2-4003.
12. Legislative Indexing Vocabulary. - Washington: Congressional Research Service. The Library of Congress., 21st Ed. 1994. - 546 p.
13. Thesaurus Internationale Beziehungen und Landerkunde.- Fachinformationverbund, 1992.
14. UNBIS Thesaurus. English Edition.- Dag Hammarskjold Library of United Nations, New York..- 1976.
15. Loukachevitch N. Knowledge Representation for Multilingual Text Categorization // Cross-Language Text and Speech Retrieval. - AAAI Symposium on Cross-Language Text and Speech Retrieval - AAAI Technical Report SS-97-05, 1997 - pp.133-142.
16. Loukachevitch N. Text Summarization Based on Thematic Representation of Texts // Intelligent Text Summarization - AAAI Symposium on Intelligent Text Summarization - AAAI Technical Report SS-98-06, 1998 - pp.77-84.
17. Loukachvitch N., Dobrov B. Construction of Structural Thematic Summary of Text. - Workshop “Text, Speech, Dialog”, - Brno, 1998 - pp.85-90.
18. Voorhees E., Harman D. Overview of the Sixth Text REtrieval Conference (TREC-6) // Information Technology: The Sixth Text REtrieval Conference (TREC-6), NIST SP 500-240, National Institute of Standards and Technology, 1998 - pp. 1-24.