ИНФОРМАТИКА

2010 **апрель-июнь** **№ 2**

УДК 004.912

Л.В. Степура

АЛГОРИТМЫ ПОСТРОЕНИЯ КОНТЕКСТА ИНФОРМАТИВНЫХ ПРЕДЛОЖЕНИЙ ПРИ РЕФЕРИРОВАНИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Рассматриваются состав и структура словарей базы знаний, используемых при реферировании текстовых документов. Предлагаются алгоритмы построения контекста информативных предложений. Для реализации данных алгоритмов формируется специальная база знаний в виде ситуативной сети.

Введение

Контекст информативного предложения, или сверхфразовое единство, – это особая синтаксико-стилистическая единица, которая примерно соответствует абзацу текста. Различие между ними состоит в том, что абзац – это единица авторского разбиения текста, а сверхфразовое единство является элементом его разбиения с учетом прагматики [1].

Абзац и межфразовое единство совпадают в тех случаях, когда изложение материала ведется по логико-смысловому принципу, т. е. если границы абзацев и межфразовых единств намечают переход от одной микротемы к другой. Такой абзац, совпадающий со сложным синтаксическим целым, называют тематическим или классическим, здесь всегда есть стержневая фраза и пояснительная часть либо имеется только стержневая фраза, которая связывает разные фрагменты текста, определяя тематический переход между ними [2]. Сверхфразовые единства всегда монотематичны, при объединении их друг с другом наблюдается переход от выражения микротем к макротеме.

При объединении предложений в сверхфразовое единство обычно используются лексические, морфологические и синтаксические признаки. К лексическим признакам относятся повторение в предложениях некоторых слов из предшествующих предложений сверхфразового единства, присутствие личных и указательных местоимений, местоименных наречий *затем*, *потом*, *тогда*, *там*, *так*. Морфологические признаки выражают одновременность или разновременность событий, описанных в объединяемых предложениях, т. е. указывают на соотношение в них видовременных форм глаголов-сказуемых. При этом используются временные союзы *когда*, *пока*, *до тех пор пока*, *лишь*, *только*, *как только*, *до того как*, *после того как*, *перед тем как*, *по мере того как*, *чуть*, *едва*, *прежде чем*, *раньше чем*, *в то время как*. Синтаксические признаки определяют порядок слов и предложений с использованием союзов типа *зато*, *однако*, *так что* и др.

В настоящей статье вместо традиционных лексических, морфологических и синтаксических признаков, используемых при построении сверхфразовых единств, предлагаются алгоритмы их формирования на основе выделения в тексте наиболее информативных предложений в качестве смысловых центров сверхфразовых единств. При поиске других предложений, образующих монотематические группы, учитываются ситуативные связи между словами и предложениями с помощью специальной базы знаний в виде ситуативной сети. Такой подход обеспечивает универсальность алгоритмов выявления монотематических фрагментов текстовых документов в смысле реферирования текстов на различных входных языках.

1. Ситуативная сеть

Процессу построения сверхфразовых единств в текстовых документах предшествует выявление в них информативных слов и предложений. Для реализации алгоритмов построения таких единств сформируем специальную базу знаний в виде ситуативной сети, т. е. графа, вершинами которого являются информативные слова, а ребрами – ситуативные связи между ними.

Основой для построения ситуативной сети являются тематические корпусы текстов [3], т. е. наборы текстов по каждой конкретной тематике предметной области. Тематические корпусы образуют полный корпус текстов. Количество полных корпусов в системе реферирования соответствует числу входных языков.

1.1. Тематические корпусы текстов

Любое непустое подмножество цепочек входного языка будем называть *текстом*, если на этом подмножестве определено отношение линейного порядка. Цепочки текста назовем *предложениями*.

Пусть имеется некоторое непустое множество текстов (совокупность текстов по конкретной тематике). Сформируем текст *Ct*, объединив все множества предложений каждого из этих текстов, и назовем его *тематическим корпусом* текстов. Поскольку в информационной системе представлено, как правило, несколько таких корпусов, будем обозначать их Ct_j (j – номер корпуса). Объединение $Cf_i = \bigcup_{j=1}^{n} Ct_j$ всех тематических корпусов назовем *полным корпусом* текстов (i – номер полного корпуса текстов).

1.2. Информативность слов и словосочетаний

Информативность слов и словосочетаний вычисляется с использованием результатов синтаксической и статистической обработки тематических корпусов текстов Ct_j и полного корпуса текстов Cf [4].

Рассмотрим следующую совокупность событий:

S_{Ct} – извлечение некоторого слова a случайным образом из тематического корпуса текстов (или текстового документа) Ct ($Ct \in Cf$);

S_{Cf} – извлечение слова a из полного корпуса текстов Cf;

H_{Ct} – появление тематического корпуса текстов (или документа) Ct.

Пусть $P(S_{Ct} / S_{Cf})$ – условная вероятность того, что слово a извлечено из множества Ct при условии, что оно уже извлечено из полного корпуса текстов Cf. Эта вероятность вычисляется следующим образом:

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct} \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_{Ct}) \cdot P(S_{Cf} / S_{Ct})}{P(S_{Cf})}.$$

Вероятность $P(S_{Ct} / S_{Cf})$ будем называть *информативностью* слова a в тематическом корпусе текстов (или текстовом документе) Ct. Слово a назовем *информативным*, если вероятность $P(S_{Ct} / S_{Cf})$ не меньше некоторого p_0 (значение $P(S_{Ct} / S_{Cf}) = p_0$ определяется эмпирически).

Условная вероятность $P(S_{Cf} / S_{Ct}) = 1$, поскольку событие, состоящее в том, что слово a извлечено из полного корпуса Cf при условии, что оно уже извлечено из тематического корпуса Ct, является достоверным, так как Ct – подмножество множества Cf. В итоге получим

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct})}{P(S_{Cf})}.$$

Вычислив $P(S_{Ct})$ по формуле полной вероятности, будем иметь

$$P(S_{Ct} / S_{Cf}) = \frac{P(S_{Ct} / H_{Ct})}{P(S_{Cf})} \cdot P(H_{Ct}).$$

При достаточно больших объемах полного корпуса текстов Cf и тематического Ct можно считать, что

$$P(S_{Ct} / H_{Ct}) \approx \frac{n_{Ct}}{N_{Ct}},\ P(S_{Cf}) \approx \frac{n_{Cf}}{N_{Cf}},\ P(H_{Ct}) \approx \frac{N_{Ct}}{N_{Cf}},$$

где n_{Ct}, n_{Cf} – абсолютные частоты встречаемости (с точностью до синонимии и словоизменения) слова a в тематическом и полном корпусах текстов; N_{Ct}, N_{Cf} – число вхождений слова a в корпусы текстов Ct и Cf соответственно. Тогда формула для вычисления информативности I^a_{Ct} слова a в тематическом корпусе текстов (или текстовом документе) Ct примет вид

$$I^a_{Ct} = \frac{n_{Ct}}{n_{Cf}}.$$

Информативность I_π предложения π или любого другого словосочетания вычисляется как длина вектора, компонентами которого являются информативности I_a, I_b, … всех словоформ a, b, … предложения, т. е.

$$I_\pi = \sqrt{I_a{}^2 + I_b{}^2 + \ldots}.$$

1.3. Словари базы знаний

При реферировании текстовых документов в системе используются следующие словари базы знаний: частотный словарь информативных словоформ, словарь синонимичных словоформ и словарь словоизменительных парадигм.

Словарь информативных словоформ. Пусть a – некоторая словоформа, P_{Cf_i} и $P_{Ct_{ij}}$ ($i = \overline{1, m}$, $j = \overline{1, n}$) – ее абсолютные частоты соответственно в i-м полном и j-м тематическом корпусах текстов. Тогда совокупность кортежей типа $\langle a, P_{Cf_i}, P_{Ct_{i1}}, P_{Ct_{i2}}, \ldots, P_{Ct_{in}} \rangle$ будем называть *словарем информативных словоформ* (табл. 1).

Таблица 1

Состав и структура словаря информативных словоформ

Словоформа	Частота в Cf_i	Частота в Ct_1	…	Частота в Ct_n	Код (номер) парадигмы
…					
играет	0095281	0038254	…	0011520	00000074
играют	0130629	0052431	…	0019263	00000074
…					
лед	0003782	0000629	…	0000491	00000125
льда	0005128	0001135	…	0000736	00000125
…					

В отличие от информативности словоформы в некотором тексте здесь под информативностью словоформы a будем понимать условную вероятность того, что эта словоформа извлечена из тематического корпуса текстов (в котором она встречается с максимальной частотой) при условии, что она уже извлечена из полного корпуса текстов. На практике, как показано выше, при достаточно больших объемах корпусов текстов информативность словоформы может быть вычислена по формуле

$$I_a = \frac{n_{\max}}{N},$$

где $n_{\max}$, N – абсолютные частоты словоформы a (с точностью до синонимии) в тематическом и полном корпусах текстов. При этом выбирается такой тематический корпус, в котором $n_{\max}$ принимает наибольшее значение по сравнению со всеми остальными корпусами. В словарь информативных словоформ включаются те из них, информативность I_a которых не меньше некоторого порогового уровня (определяется эмпирически).

В словаре информативных словоформ каждой словоформе поставлены в соответствие:
– частота в полном корпусе текстов;
– частоты в тематических корпусах текстов;
– номер (код) парадигмы. В первоначальном состоянии каждая словоформа словаря образует отдельную парадигму. После объединения словоформ в словоизменительные парадигмы словоформам присваивается номер парадигмы, элементом которой эта словоформа является.

Словарь синонимичных словоформ состоит из групп синонимичных словоформ, которые могут быть использованы при определении их информативности (табл. 2).

Таблица 2

Состав и структура словаря синонимичных словоформ

Словоформа	Синонимичные словоформы
...	
языкознание	лингвистика
	языковедение
...	

Словарь синонимичных словоформ создается «вручную» с использованием средств визуализации автоматизированного рабочего места (АРМ) эксперта-лингвиста.

Словарь словоизменительных парадигм служит для поиска всех словоформ парадигмы после нахождения словоформы и ее кода в словаре словоформ. Процедура такого поиска используется при вычислении информативности слов. Словарь парадигм создается и актуализируется в человеко-машинном режиме с использованием соответствующего инструментария АРМ эксперта-лингвиста. В первоначальном состоянии каждая парадигма словаря парадигм содержит одну-единственную словоформу для каждого кода словоформы. После формирования парадигм коды меняются (табл. 3 и 4).

Таблица 3

Состав и структура словаря парадигм (промежуточное состояние)

Код (номер) парадигмы	Словоформа
...	
00000074	играет
00000075	играют
...	
00000125	лед
00000126	льда
00000127	льдом
...	

Таблица 4

Состав и структура словаря парадигм (конечное состояние)

Код (номер) парадигмы	Парадигма
...	
00000074	играет
	играют
...	
00000125	лед
	льда
	льдом
...	

1.4. Алгоритм формирования словаря информативных словоформ

Пусть W_{Cf_i} – множество всех словоформ полного корпуса текстов Cf_i. Представим словарь информативных словоформ в виде $Inf = \{Inf_c = \langle c, I_c \rangle | c \in W_{Cf_i}\}$. Сформируем словарь Inf по следующему алгоритму.

Алгоритм 1. На входе алгоритма словарь словоформ W (словарь включает все словоформы полного корпуса текстов; его структура аналогична структуре словаря информативных словоформ, представленного табл. 1), на выходе – словарь информативных словоформ Inf. Алгоритм состоит из следующих шагов:

1. $Inf := \varnothing$, $i := 1$.

2. Выбрать из словаря словоформ W кортеж $W_a = \langle a, P_{Cf_i}, P_{Ct_{i1}}, P_{Ct_{i2}}, \ldots \rangle$ $\langle a, P_{Cf_i}, P_{Ct_{i1}}, P_{Ct_{i2}}, \ldots, P_{Ct_{in}} \rangle$.

3. Найти в кортеже W_a максимальную частоту $P_{Ct_{ij}}$, с которой словоформа *a* входит в тематический корпус текстов.

4. $I_a := P_{Ct_{ij}} / P_{Cf_i}$.

5. Поместить кортеж $Inf_a = \langle a, I_a \rangle$ в словарь *Inf*.

6. Если элементы словаря *W* исчерпаны, то конец (словарь информативных словоформ сформирован). Иначе $i := i + 1$, перейти к п. 2.

1.5. Ситуативное отношение. Определение ситуативной сети

Пусть Ct_j ($j = \overline{1, n}$; $n \geq 2$) – тематические корпусы текстов, Cf ($Cf = Ct_1 \cup Ct_2 \cup Ct_n$) – полный корпус текстов, а *Wo* – множество всех слов полного корпуса текстов *Cf*. Введем в рассмотрение для каждого полного корпуса текстов *Cf* ситуативное отношение.

Отношение толерантности Θ (рефлексивное и симметричное бинарное отношение) на множестве *Wo* назовем *ситуативным отношением* в полном корпусе текстов *Cf*, если любая упорядоченная пара информативных слов (*a*, *b*) из множества *Wo* является элементом отношения Θ тогда и только тогда, когда вероятность совместной встречаемости слов *a* и *b* в корпусе текстов *Cf* не меньше некоторого порогового значения. (Эту вероятность будем называть *информативностью ситуативной связи* слов.) Под совместной встречаемостью двух слов здесь понимается их наличие (а также их синонимов и словоизменений) в одном и том же предложении корпуса *Cf*. Граф $S_{\text{сит.}}$ ситуативного отношения, каждое ребро которого помечено значением информативности соответствующей ситуативной связи, будем называть *ситуативной сетью* (рис. 1).

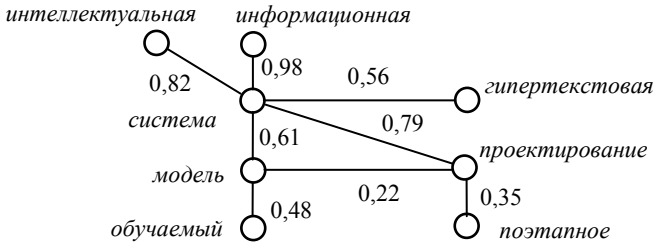

Рис. 1. Фрагмент ситуативной сети

1.6. Ситуативный словарь

На практике сеть $S_{\text{сит.}}$ удобно представлять в виде ситуативного словаря (табл. 5). В первых двух столбцах табл. 5 представлены пары слов, а в третьем – информативность ситуативной связи этих слов в процентах.

Таблица 5

Фрагмент ситуативного словаря

Слово	Слово	Информативность ситуативной связи слов, %
...		
система	интеллектуальная	82
система	информационная	98
система	гипертекстовая	56
модель	система	61
модель	обучаемый	48
проектирование	система	79
проектирование	модель	22
проектирование	поэтапное	35
...		

1.7. Информативность ситуативных связей между предложениями текста

Пусть π и ρ – произвольные предложения или словосочетания. Обозначим через I_{ab} информативность ситуативной связи произвольных информативных слов *a* и *b*. Тогда информативность ситуативных связей между предложениями π и ρ будем вычислять по аналогии с вычислением информативности предложений по формуле

$$I_{\pi\rho} = \sqrt{I_{ab}{}^2 + I_{cd}{}^2 + \ldots},$$

где *a*, *c*, … – слова предложения π; *b*, *d*, … – слова предложения ρ.

2. Построение сверхфразовых единств

При построении в тексте сверхфразовых единств вначале будем использовать ситуативные связи между словами текста, а затем между его предложениями. Ситуативные связи между предложениями текста представим в виде графа ситуативных связей.

2.1. Граф информативности текста

Пусть имеется текст (т. е. кортеж предложений) $T = \langle \pi_1, \pi_2, \ldots \rangle$. Вычислим информативность всех предложений текста *T* и исключим из *T* неинформативные предложения, т. е. все предложения π, информативность I_π которых меньше некоторого I_0. В результате получим кортеж предложений $T_{\text{инф.}} = \langle \pi_{i_1}, \pi_{i_2}, \ldots \rangle$, который будем называть *маршрутом информативности* текста *T*. Соединив последовательно вершины графа текста G_T (т. е. графа редукции линейного порядка на множестве всех предложений текста *T*), соответствующие информативным предложениям, получим орграф $G_{\text{инф.}}$, который будем называть *графом информативности* текста *T* (рис. 2). Вершины и дуги графа текста, не вошедшие в состав графа информативности $G_{\text{инф.}}$, изображены пунктирными линиями.

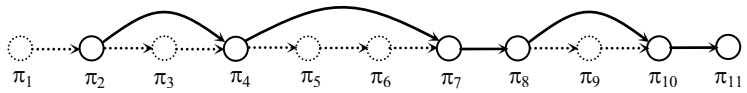

Рис. 2. Пример графа информативности текста

2.2. Граф ситуативных связей между предложениями текста

Пусть T^+ – множество всех информативных, а T^- – всех неинформативных предложений текста *T* ($T = T^+ \cup T^-$). Определим на паре множеств T^+, T^- симметричное отношение Ξ, такое, что для любых предложений $\pi \in T^+$ и $\rho \in T^-$ $(\pi, \rho) \in \Xi$ тогда и только тогда, когда информативность $I_{\pi\rho}$ ситуативной связи между предложениями π и ρ не меньше некоторого значения. Граф отношения Ξ назовем *графом ситуативных связей* между предложениями текста *T* (рис. 3). Информативные предложения изображены на рис. 3 сплошными линиями, а неинформативные – пунктирными. Пунктирными стрелками представлены дуги графа текста, а скобками объединены возможные «кандидаты» в сверхфразовые единства.

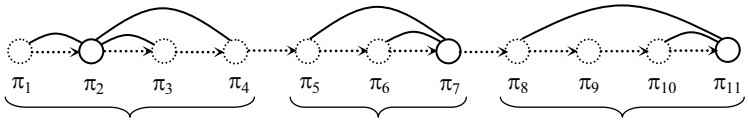

Рис. 3. Пример графа ситуативных связей между предложениями текста

2.3. Алгоритмы разбиения текста на сверхфразовые единства

Процесс разбиения текста на сферхфразовые единства осуществляется двумя алгоритмами. В соответствии с первым алгоритмом в тексте выявляются информативные предложения, т. е. строится маршрут информативности $T_{\text{инф.}}$. Второй алгоритм предназначен для установления ситуативных связей между предложениями, т. е. для создания сверхфразовых единств.

Алгоритм 2. На входе алгоритма – текст T объемом в N предложений, на выходе – текст T с выделенными информативными предложениями. Алгоритм состоит из следующих шагов:

1. Вычислить информативность I_a каждой словоформы a текста T по формуле

$$I_a = \frac{n_T}{n_{Cf}},$$

где n_T – абсолютная частота встречаемости словоформы a в тексте T.

2. Вычислить информативность I_π каждого предложения π текста T по формуле

$$I_\pi = \sqrt{I_a{}^2 + I_b{}^2 + \ldots},$$

где I_a, I_b,… – значения информативности всех слов предложения π.

3. Упорядочить все N предложений текста T по убыванию их информативности.

4. Выделить из полученного упорядоченного списка n наиболее информативных предложений (n задается в качестве параметра).

5. Восстановить исходный порядок предложений в тексте T. Конец.

Алгоритм 3. На входе алгоритма – текст T с выделенными информативными предложениями, на выходе – сверхфразовые единства $E_1, E_2, \ldots, E_n$. Алгоритм состоит из следующих шагов:

1. $E_i := \varnothing,\ i = \overline{1, n}$ (i – номер информативного предложения текста T).

2. $i := 1$.

3. $\pi := \pi_i$.

4. Вычислить значения информативности ситуативных связей $I_{\pi\sigma}$ для всех неинформативных предложений σ, предшествующих предложению π и следующих за ним, по формуле

$$I_{\pi\sigma} = \sqrt{I_{ab}{}^2 + I_{cd}{}^2 + \ldots}.$$

5. Поместить все предложения σ, для которых $I_{\pi\sigma}$ превышает пороговый уровень $I^0_{\pi\sigma}$ (задается в качестве параметра), в множество E_i (в порядке их появления в тексте T).

6. $i := i + 1$. Если $i \le n$, то перейти к п. 4, иначе – конец (сверхфразовые единства построены).

Пример построения сверхфразовых единств. В соответствии с алгоритмом 2 в исходном тексте [5] выявлены следующие информативные предложения:

1. Через два месяца Поль объявился неподалеку, в курортном местечке Гузе-Нейж.

2. Кроме того, он рассказывал, что долгие годы прослужил священником в храме, молясь вместе с прихожанами и прося Бога, чтобы он избавил их от злобы волков с женскими головами.

3. Разумеется, многие сочли геолога сумасшедшим, но некоторые, вспоминая рассказы стариков, верили ему.

По алгоритму 3 сформированы сверхфразовые единства:

1. В сентябре 1998 года молодой геолог Поль Леблан, отстав от группы, заблудился в долине Тургвилла неподалеку от озера Алет. Его встревоженные товарищи искали Поля всю ночь, а потом еще несколько дней, просматривая буквально каждый квадратный метр местности с вертолета, но все безрезультатно.Через два месяца Поль объявился неподалеку, в курортном местечке Гузе-Нейж.

2. Местные жители были шокированы его рассказами и вопросами. Геолог утверждал, что ему 33 года, хотя выглядел он на 50. Кроме того, он рассказывал, что долгие годы прослужил священником в храме, молясь вместе с прихожанами и прося Бога, чтобы он избавил их от злобы волков с женскими головами. Поль хорошо помнил, что когда-то был геологом, но утверждал, что было это много лет назад.

3. Разумеется, многие сочли геолога сумасшедшим, но некоторые, вспоминая рассказы стариков, верили ему.

Заключение

При реферировании текстовых документов выделяются информативные предложения и с их помощью формируется реферат. В статье предложена модель, улучшающая качество реферата за счет выделения сверхфразовых единств, смысловыми центрами которых являются информативные предложения, попавшие первоначально в реферат. На основе модели разработаны алгоритмы построения сверхфразовых единств. В дальнейшем при реферировании каждое информативное предложение будет заменено на сверхфразовое единство и реферат будет представлен в расширенном варианте.

Список литературы

1. Розенталь, Д.Э. Словарь лингвистических терминов / Д.Э. Розенталь [Электронный ресурс]. – Режим доступа : http://www.classes.ru/grammar/114.Rosental/17-s-2/html/unnamed_36.html. – Дата доступа : 14.01.2010.

2. Валгина, Н.С. Теория текста : учеб. пособие / Н.С. Валгина. – М. : Мир книги, 1998. – 210 с.

3. Липницкий, С.Ф. Индексирование и реферирование текстовых документов по космической тематике / С.Ф. Липницкий, Л.В. Степура // Материалы Четвертого Белорусского космического конгресса, Минск, Республика Беларусь, 27–29 октября 2009 г. – Минск : ОИПИ НАН Беларуси, 2009. – С. 185–190.

4. Кравцов, А.А. Система автоматического индексирования и реферирования текстовых документов / А.А. Кравцов, С.Ф. Липницкий, Л.В. Степура // Таврический вестник информатики и математики. – 2008. – № 1. – С. 260–266.

5. Вход в параллельный мир [Электронный ресурс]. – Режим доступа : http://emigration.russie.ru/news/8/11248_1.html. – Дата доступа : 14.01.2010.

Поступила 15.01.10

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: stepura@newman.bas-net.by

L.V. Stepura

THE ALGORITHMS OF CONTEXT CREATION OF THE INFORMATIVE SENTENCES FOR ABSTRACTING TEXT DOCUMENTS

The composition and structure of the dictionary knowledge base used for abstracting text documents are considered. The algorithms of context creation of the informative sentences are suggested. For implementation of the given algorithms the special knowledge base as a situational network is generated.