4
П-71
36

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР

Проблемная группа по экспериментальной
и прикладной лингвистике

ПРЕДВАРИТЕЛЬНЫЕ ПУБЛИКАЦИИ

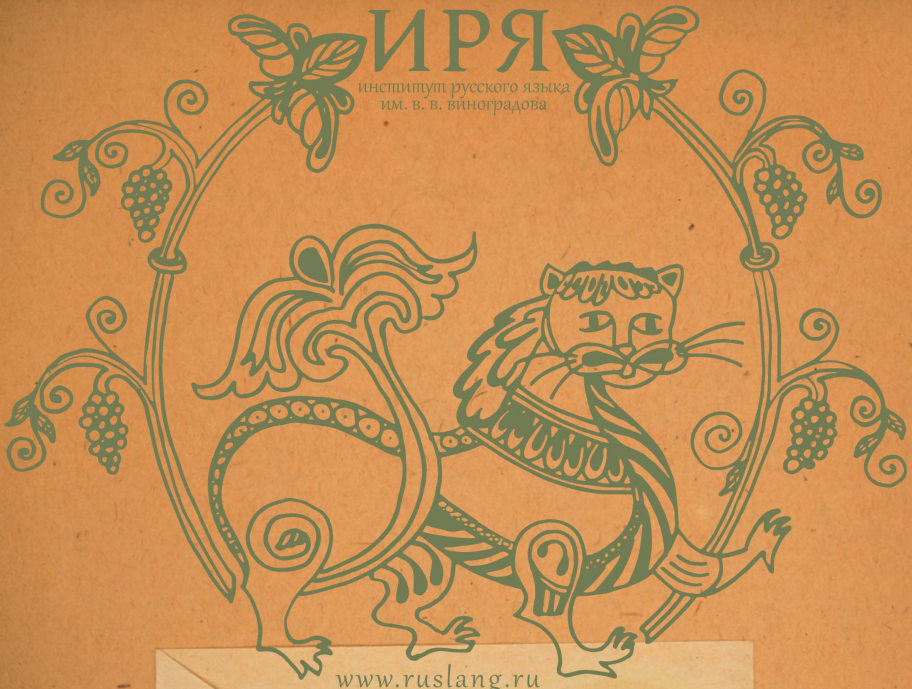
Выпуск 36

Н. Н. ЛЕОНТЬЕВА, Е. В. УРЫСОН

АЛГОРИТМ ПОСТРОЕНИЯ ИНФОРМАЦИОННОЙ ЗАПИСИ ДЛЯ ТЕКСТА (I ЭТАП)

Москва—1973

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

67263

ИНСТИТУТ РУССКОГО ЯЗЫКА АН СССР
Проблемная группа по экспериментальной
и прикладной лингвистике

Предварительные публикации

Выпуск 36

Н.Н. Леонтьева, Е.В. Урысон

АЛГОРИТМ ПОСТРОЕНИЯ ИНФОРМАЦИОННОЙ ЗАПИСИ ДЛЯ ТЕКСТА (I ЭТАП)

Москва - 1973

Аннотация.

Описывается алгоритм автоматического перевода фразы с русского языка на информационный язык (ИЯ). В качестве ИЯ используется искусственный язык, приближающийся по своим семантическим возможностям к естественному языку (ЕЯ). Перевод некоторого целого текста с ЕЯ на ИЯ осуществляется в несколько последовательных преобразований (этапов). Первым этапом является перевод каждой отдельной фразы с ЕЯ на ИЯ (называемый естественно-семантическим анализом фразы, или ЕС-анализом). В статье подробно описан этап ЕС-анализа, основная задача которого - установление соответствий между единицами ЕЯ и ИЯ, приведены результаты первого машинного эксперимента по ЕС-анализу. Кратко описано содержание следующих этапов преобразований, конечной целью которых является построение информационной записи целого текста.

4
П-71136

БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

67263

0. Введение

Информационной записью (ИЗ) текста мы называем такое выражение на информационном языке (ИЯ), в котором сохранена вся та смысловая информация исходного текста, которая необходима для решения информационных задач определенного типа. В частности, поисковый образ документа (ПОД) можно считать ИЗ данного документа с точки зрения задач только документального поиска. С точки зрения задачи фактографического поиска или задачи построения ответов на вопросы по данному тексту ПОД уже не может считаться его ИЗ. Задачи большей информационной сложности, чем документальный поиск, предъявляют вполне определенные требования к ИЗ и тем самым к структуре ИЯ: это должен быть язык, приближающийся по своим семантическим возможностям к естественному языку, но обладающий гораздо более простой структурой. Последнее необходимо для того, чтобы можно было выражение на ИЯ подвергать разного рода формальным преобразованиям (сопоставление, извлечение необходимого фрагмента информации, классификация, сжатие и др.). Термин ИЗ мы оставим только для окончательного выражения на ИЯ, соответствующего смыслу целого текста; любое промежуточное выражение на ИЯ, соответствующее предложению или тексту, будем называть семантической записью (СЗ) этого предложения или текста.

Построение СЗ отдельной, т.е. "вынутой из контекста", фразы - первый этап создания ИЗ любого текста. Этот этап называется естественно-семантическим анализом фразы (ЕС-анализом); СЗ, полученная в результате ЕС-анализа, называется первичной СЗ.

В данной статье рассматривается алгоритм построения первичной СЗ отдельной фразы, или алгоритм ЕС-анализа. В качестве ИЯ мы использовали искусственный язык, названный условно ЭХО [3]. Язык ЭХО создавался с установкой на решение сложных информационных задач, в первую очередь фактографического поиска и автоматического реферирования.

Ниже перечисляются такие свойства и компоненты ИЯ ЭХО, которые необходимы для описания работы алгоритма ЕС-анализа.

Алгоритм ЕС-анализа работает после морфологического анализа. Отдельного этапа синтаксического анализа в системе нет, т.е. не строится синтаксическое дерево предложения с так называемыми поверхностно-синтаксическими связями между словами. Но отсутствие отдельного уровня синтаксического анализа не означает, что его вообще нет. ЕС-анализ является синтаксико-семантическим как по самой процедуре (при образовании формул СЗ работают и семантические, и синтаксические фильтры), так и по результатам (смысловые отношения, устанавливаемые между переводами на ИЯ слов естественного языка (ЕЯ), чаще всего являются смысловой интерпретацией синтаксических связей между соответствующими словами).

Совмещение синтаксического и семантического уровней дает существенное уменьшение количества вариантов при установке на многовариантность анализа.

1. Единицы искусственного языка

Единицами семантического языка ЭХО являются: 1) термы (дескрипторы); 2) двуместные отношения R(,) , называемые смысловыми отношениями. Множество термов является открытым, множество смысловых отношений - закрытым. Язык ЭХО располагает правилами образования формул СЗ и правилами преобразования СЗ текста.

1.1. Дескрипторы (термы). Каждый дескриптор является представителем одного слова или группы слов ЕЯ, имеющих один и тот же смысл. Дескриптор также может представлять значение, выраженное аффиксом; таковы некоторые грамматические значения, напр., "число", "вид", "время" и др. (как правило, значение этих категорий выражается дескриптором вместе со смысловым отношением, т.е. фрагментом СЗ - см. ниже). Ниже дескрипторы обозначаются словами ЕЯ, подчеркнутыми прямой чертой.

Каждый дескриптор имеет внутреннюю смысловую структуру, которая описывается семантическими признаками и смысловыми отношениями R(,), установленными на их множестве (ниже семантические признаки обозначаются словами ЕЯ, взятыми в кавычки). В экспериментальном варианте системы оказалось возможным описывать смысловую структуру дескриптора в упрощенном виде - всегда одним дескриптором.

Семантические признаки частично упорядочены между собой отношением включения. В настоящем варианте системы используется 13 семантических признаков:

1. "действие" (например, строить, внедрять)
2. "предмет" (например, камень, машина)
3. "устройство" $\subset$ "предмет" (например, экскаватор)
4. "деталь" $\subset$ "предмет" (например, кулачок)
5. "вещь" $\subset$ "предмет" (например, камень)
6. "материал" $\subset$ "предмет" (например, бетон)
7. "пространство" $\subset$ "предмет" (например, море)
8. "учреждение" $\subset$ "пространство" (например, НИИ)
9. "документ" $\subset$ "предмет" (например, справка)
10. "собирательность" (например, совокупность)
11. "абстрактность" (например, теория)
12. "область деятельности" (например, наука)
13. "характеристика" (например, синий)

В качестве семантического признака может быть также использован любой член списка отношений.

1.2. Смысловые отношения. Смысловые отношения R(,) (ниже обозначаются словами ЕЯ, подчеркнутыми пунктиром) могут устанавливаться на множестве дескрипторов и на множестве формул.

Любое высказывание ЕЯ переводится на ЭХО как последовательность формул $R(\alpha, \beta)$, где α и β - соответственно дескрипторы A_0, B_0: $R(A_0, B_0)$; α и/или β может быть более чем одноэлементным множеством дескрипторов: $\alpha = A_0, A_1, A_2, \ldots, A_n$, $\beta = B_0, B_1, B_2, \ldots B_m$, где $n>1$ или $m>1$: $R(\{A_0, A_1, \ldots, A_n\}, \{B_0, B_1, \ldots, B_m\})$; наконец, α и/или β мо-

жет быть формулой одного из этих двух видов, и вся формула может иметь такой, напр., вид: $R^n[R^i(A_c, B_c), R^j(\{A_1, A_2, \ldots, A_k\}, B_m)]$ (в экспериментальном варианте системы подобных формул нет). Одно из мест отношения $R(\alpha, \beta)$ в СЗ может быть пустым.

Приведем для наглядности пример СЗ одного предложения.

Просьба об ускорении создания центров сбора и обработки научно-технической информации.

1. субъект (, просить)
2. содержание (ускорить, просить)
3. объект (сделать, ускорить)
4. объект (учреждение, сделать)
5. специализация ({собирать, обрабатывать}, учреждение)
6. объект (информация, {собирать, обрабатывать})
7. сфера ({наука, техника}, информация)

Как было сказано выше, множество смысловых отношений является конечным. Предполагается, что для описания текстов разной тематики требуется 40–100 отношений. Для описания текстов одной и той же тематики, по-видимому, достаточно 40–50 отношений. В описываемом эксперименте использовались следующие смысловые отношения (которые безусловно необходимы для описания текстов любой отрасли):

1. связано с (,)
2. принадлежность (,)
3. часть (,)
4. вид (,)
5. параметр (,)
6. количество (,)
7. имя (,)
8. время (,)
9. локализация (,)
10. начальный пункт (,)
11. конечный пункт (,)
12. причина (,)
13. назначение (,)
14. модальность (,)
15. специализация (,)
16. содержание (,)
17. источник (,)
18. цель (,)
19. оценка (,)
20. равно (,)
21. больше (,)
22. субъект (,)
23. объект (,)
24. адресат (,)
25. с помощью (,)
26. материал (,)

Каждое смысловое отношение $R^i(\ ,\ )$ связано хотя бы с одним другим отношением $R^j(\ ,\ )$ отношением включения. Если $R^i(\alpha,\beta) \subset R^j(\alpha,\beta)$, то при преобразовании СЗ формула $R^i(\alpha,\beta)$ может быть при любых условиях заменена на формулу $R^j(\alpha,\beta)$, а формула $R^j(\alpha,\beta)$ при определенных условиях может быть заменена на формулу $R^i(\alpha,\beta)$.

Пример:

актант (В, А) ⊃ | субъект (В, А)
| объект (В, А)
| адресат (В, А)
| с помощью (А, В)

Некоторые смысловые отношения симметричны (например, отношение равно (,)). Все симметричные, а также некоторые другие отношения транзитивны, например:

локализация (А, В)
&
локализация (В, С) → локализация (А, С)

Эти свойства отношений служат средством преобразования и варьирования СЗ.

Как было сказано выше, смысловое отношение может быть использовано в качестве семантического признака. Приведем пример: слово два переводится дескриптором два, имеющим семантический признак "количество (,)".

1.3. Правила образования формул. Для каждого смыслового отношения указываются семантические признаки тех дескрипторов, которые могут заполнить места этого отношения. Например, второе место отношений субъект (,), адресат (,) может быть заполнено только таким дескриптором, который имеет семантический признак "действие" или "документ" (может быть заполнено только таким более чем одноэлементным множеством дескрипторов, что каждый из дескрипторов имеет семантический признак "действие" или "документ").

Если о двух семантических признаках "α" и "β" известно, что "α" $\subset$ "β", и об отношении $R^i(,)$ известно, что какое-либо его место может быть заполнено дескриптором с семан-

тическим признаком "β", то это же место отношения $R^i(\ ,\ )$ может быть заполнено дескриптором с семантическим признаком "α". Например, первое место отношения локализация (,) может быть заполнено дескриптором с семантическим признаком "пространство", а второе - дескриптором с семантическим признаком "предмет" или "действие". Известно, что "учреждение" $\subset$ "пространство", поэтому правильной является как формула локализация (пустыня, исследовать), так и формула локализация (НИИ, исследовать) (пустыня - "пространство", НИИ - "учреждение").

Мы не будем приводить здесь списка разрешаемых семантических признаков для каждого места каждого отношения. Ниже мы еще раз вернемся к вопросу образования правильных формул.

2. Структура естественно-семантического словаря (ЕС-словаря)

ЕС-словарем называется "ЕЯ - ЭХО словарь", т.е. словарь, который единицам естественного (русского) языка ставит в соответствие единицы ИЯ ЭХО. Словарная статья "ЕЯ - ЭХО словаря" состоит из нескольких частей (граф). Рассмотрим эти части.

2.1. Графа "Перевод слова". 1) Большинство слов ЕЯ переводятся дескрипторами, например, строить - строить, испытания - проверять, работа$_1$ - работать, работа$_2$ - публикация, новый - новый и т.д. В качестве дескрипторов, как было сказано выше, используются слова ЕЯ, а именно, глаголы, существительные и прилагательные в так называемой словарной форме (т.е. глаголы в форме инфинитива, существительные - в форме именительного падежа единственного числа, прилагательные - в форме мужского рода именительного падежа единственного числа). При этом, если дескриптор имеет семантический признак "действие", то для обозначения дескриптора выбирается глагол (и обратно); если дескриптор имеет семантический признак "абстрактность" или "область деятельности" или "собирательность" или "предмет" или семантический признак "α" такой, что "α" $\supseteq$ "предмет", то для обозначения дескриптора выбирается существительное (и об-

ратно); если дескриптор имеет семантический признак "характеристика", то для обозначения дескриптора выбирается прилагательное (и обратно).

2) Некоторые слова ЕЯ переводятся на ЭХО не дескрипторами, а отношениями. Таковы, например, слова равно (равно (,)), больше (больше (,)); таковы также предлоги в некоторых фразах, напр.: дом у дороги (локализация (,)), строили в апреле (время (,)).

3) Некоторые слова (в основном, сложные) переводятся несколькими дескрипторами, между которыми устанавливаются смысловые отношения. Напр., слово снегоуборочный переводится двумя дескрипторами: снег и убирать, между которыми установлено следующее отношение: объект (снег, убирать). Слово небольшой переводится дескрипторами не и большой, между которыми установлено отношение модальность (не, большой).

Если слово переводится несколькими дескрипторами, то один из дескрипторов выбирается в качестве "основного". Например, "основным" дескриптором для слова снегоуборочный является дескриптор убирать, для слова небольшой - дескриптор большой. В дальнейшем, говоря, что слово $\underset{\sim}{a}$ в данном значении переводится дескриптором A, мы будем подразумевать следующее: A является "основным" или единственным дескриптором, которым переводится слово $\underset{\sim}{a}$ в данном значении. В графе "Перевод слова" записывается только "основной" дескриптор; все прочие дескрипторы попадают в графу "Набор валентностей" (см. ниже).

4) В экспериментальном варианте системы любое слово (в одном из своих значений) переводится либо дескриптором, либо отношением, либо несколькими дескрипторами. Но, вообще говоря, некоторые слова можно переводить множеством дескрипторов и отношений, на котором устанавливаются отношения. Например, слово столы можно переводить дескрипторами стол, один, причем дескриптор один входит в формулу: больше (, один); эта формула и дескриптор стол заполняют места отношения количество (,): количество (больше (, один), стол) (различием

между формами единственного и множественного числа существительных можно во многих случаях пренебрегать, т.е. не отражать это различие в СЗ).

2.2. Графа "Семантический признак перевода". Как уже было сказано выше, дескриптор имеет семантическую структуру, которая описывается семантическим признаком. Семантический признак дескриптора помещается в специальной графе словарной статьи. Если слово переводится отношением, то эта графа остается пустой. Список семантических признаков с примерами см. выше.

2.3. Графа "Набор валентностей". С точки зрения перевода текста с ЕЯ на ЭХО очень важным является понятие смысловой валентности. Смысловая валентность - это словарная характеристика такого слова ЕЯ, которое на язык ЭХО переводится дескриптором. Смысловая валентность слова α, переводимого дескриптором А, - это такое смысловое отношение $R^i(\ ,\ )$, что дескриптор А заполняет одно из его мест и во многих фразах с данным словом α имеется слово β, переводимое дескриптором В, который заполняет другое место этого отношения. Например, слово перевозка, переводимое дескриптором перевозить, имеет следующие валентности: субъект (\ \ , перевозить), объект (\ \ , перевозить), исходный пункт (\ \ , перевозить), конечный пункт (\ \ , перевозить), с помощью (перевозить, \ \ ).

Смысловые валентности в нашей системе привязаны именно к словам ЕЯ, а не к дескрипторам. Существуют такие слова ЕЯ, которые переводятся одним и тем же дескриптором, но имеют разные валентности. Например, слова река и речной переводятся одним и тем же дескриптором река, но у первого из этих слов есть валентность имя (\ \ , река) (ср. река Ока), а у второго - валентность принадлежность (\ \ , река) (ср. речная вода, рыба, речной песок).

Слово может не иметь ни одной валентности (например, не имеют валентностей так называемые предметные существительные: снег, стол и т.п.); максимальное количество валентностей, предусмотренное в ЕС-словаре, равно шести (т.е. графа "Набор валентностей" разделена на 6 частей).

Перевод слова α в СЗ появляется не меньшее число раз, чем число валентностей: в СЗ записываются все валентности слова α (в том числе и незаполненные).

Если слово переводится отношением, то графа "Набор валентностей" не заполняется.

2.4. Графа "Семантические требования (СТ) валентности (места)". Для каждой валентности каждого слова указывается, дескриптор с каким семантическим признаком может ее заполнить (эта информация называется семантическим требованием (СТ) валентности). Например, валентность объект (, писать) может быть заполнена дескриптором с семантическим признаком "документ" и только им.

Заметим, что семантический признак дескриптора в нашей системе не полностью отражает смысл слова, поэтому, например, валентности объект (, писать) и объект (, строить) имеют разные СТ - несмотря на то, что дескрипторы писать и строить имеют один и тот же семантический признак "действие"; а именно, первая валентность имеет СТ "документ", а вторая - СТ "устройство" или "деталь" или "вещь" или учреждение"[1].

Если слово переводится отношением $R(,)$, то в этой графе для каждого его места указываются семантические признаки тех дескрипторов, которые могут его заполнить.

2.5. Графа "Синтаксические условия заполнения валентностей (мест)". Для каждой валентности (места) указывается не только информация о дескрипторе X, могущем заполнить эту валентность (место), но и информация о слове α, переводом которого является дескриптор X. Эта информация записывается в графу "Синтаксические условия заполнения валентностей (мест)". Графа состоит из двух частей. В первой части указывается часть речи, к которой принадлежит слово α, падеж или падеж и число - если α - существительное; род, число и падеж - если α - прилага-

[1] Ср. интересную статью [6], в которой предлагается аппарат получения набора валентностей слова из набора его семантических признаков.

тельное или причастие (эта часть информации называется морфологическим требованием (МТ) валентности). Во второй части графы указывается позиция, которую занимает во фразе слово x относительно описываемого слова a ; имеются 4 позиции: справа, слева, непосредственно справа, непосредственно слева (для анализируемых текстов самой частой является первая позиция). Пример: валентность слова писать объект (, писать) может быть заполнена таким дескриптором с семантическим признаком "документ", что этот дескриптор является переводом существительного в форме винительного падежа, причем это существительное находится во фразе правее слова писать (ср. писать рекомендацию).

Такое более или менее "жесткое" задание позиции весьма эффективно для анализа кратких резюме и заголовков; но, вообще говоря, для других текстов оно не годится и, в принципе, можно обойтись заданием позиции для весьма небольшого круга слов (например, для прилагательных и предлогов).

Если валентность слова может быть заполнена таким дескриптором, что во фразе на ЕЯ слово, переводом которого является этот дескриптор, обязательно управляется предлогом (ср., верить в победу, специалист по экскаваторам, работа о подшипниках), то предлог записывается в МТ валентности вместе с управляемым падежом.

У некоторых валентностей (мест) есть более чем одно СТ и более чем одно МТ, причем определенные СТ "привязаны" к определенным МТ, т.е. одна и та же валентность (место) может быть заполнена дескриптором X с семантическим признаком "x" при соблюдении одних синтаксических условий или дескриптором Y с семантическим признаком "y" - при соблюдении других синтаксических условий (примеры см. ниже). Эта ситуация также отражается в ЕС-словаре.

2.6. Графа "Несовместимые валентности". В этой графе указывается информация о несовместимости заполнения некоторых валентностей одного и того же слова или о несовместимости условий их заполнения (в описываемом экспериментальном варианте ЕС-анализа эта информация не использовалась).

2.7. Приведем в качестве иллюстрации несколько словарных статей.

Завод

Дескриптор:	Семантический признак:	Валентности:	СТ	МТ	Позиция во фразе
завод	"учреждение"	1. специализация(,завод)	"предмет"	сущ.род.	справа
			"действие"	по+сущ.дат.; для+сущ.род.	
		2. имя (,завод)	любой	имени+сущ.род.	справа
			"имя(,)"	сущ. им. в кавычках	

Бетоносмесительный

Дескриптор:	Семантический признак:	Валентности:	СТ	МТ	Позиция во фразе
смешивать	"действие"	1. объект(бетон, смешивать)			
		2. специализация(смешивать,)	"устройство"	сущ.соглас.	справа

Выгрузка

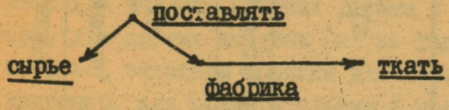
Дескриптор:	Семантический признак:	Валентности:	СТ	МТ	Позиция во фразе
выгружать	"действие"	1. объект(,выгружать)	"предмет"	сущ. род.	справа
		2. начальный пункт(,выгружать)	"предмет"	из+сущ.род.	справа
		3. конечный пункт(,выгружать)	"предмет"	на+сущ.вин.; в+сущ.вин.	справа

3. Структура первичной СЗ.

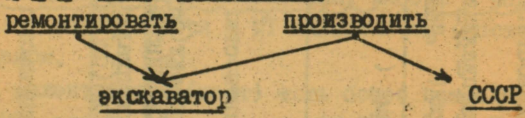
Первичная СЗ фразы представляет собой граф, причем граф упорядоченный (все слова фразы перенумерованы и перевод каждого слова a_i имеет тот же порядковый номер, что и само слово a_i). Этот граф является связным. В отличие от графа, характеризующего синтаксическую структуру фразы, первичная СЗ не обязана быть деревом. Рассмотрим примеры.

Поставка сырья ткацкой фабрике
объект (сырье, поставлять)
адресат (фабрика, поставлять)
специализация (ткать, фабрика)

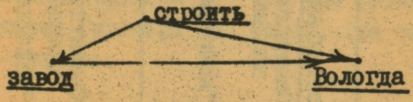
Для наглядности можно обозначить каждое отношение стрелкой (направленной от второго члена отношения к первому):



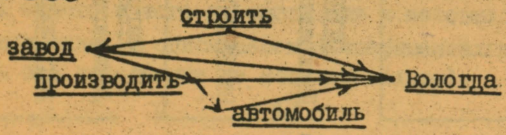
Ремонт экскаваторов отечественного производства
объект (экскаватор, ремонтировать)
объект (экскаватор, производить)
субъект (СССР, производить)



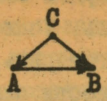
Строительство завода в Вологде
объект (завод, строить)
локализация (Вологда, завод)
локализация (Вологда, строить)



В Вологде построили завод по производству автомобилей



Ситуации, когда в СЗ фразы допустима "встреча стрелок", пока не исследованы. Подграфы вида



возможны в СЗ некоторых фраз со слабыми предложными связями (см. ниже). Ситуации, когда в СЗ фразы допустимы подобные подграфы, рассмотрены в работе [4].

Если во фразе есть однородные члены, то в СЗ фразы есть хотя бы одна формула вида $R^i(\{A_1, A_2, \ldots, A_n\}, \{B_1, B_2, \ldots, B_m\})$, где $n>1$ или $m>1$ [2].

Пример. Поставка сырья и новых станков ткацкой фабрике

объект ({сырье, станок}, поставлять)
оценка (новый, станок)
адресат (фабрика, поставлять)
специализация (ткать, фабрика).

4. Описание алгоритма ЕС-анализа.

Для краткости, вместо того, чтобы говорить: "Дескриптор А (отношение $R^n(,)$), который (которое) является переводом слова $\underset{\sim}{a}$, и дескриптор В (отношение $R^m(,)$), который (которое) является переводом слова $\underset{\sim}{b}$, заполняют такие-то места отношения $R^i(,)$", будем говорить: "Слова $\underset{\sim}{a}$ и $\underset{\sim}{b}$ заполняют такие-то места отношения $R^i(,)$"; аналогично, вместо того чтобы говорить: "Дескриптор В, который является переводом слова $\underset{\sim}{b}$, заполняет такую-то валентность слова $\underset{\sim}{a}$", будем говорить: "Слово $\underset{\sim}{b}$ заполняет такую-то валентность слова $\underset{\sim}{a}$"; наконец, вместо того чтобы говорить: "Дескриптор В (отношение $R^i(,)$), который (которое) является переводом слова $\underset{\sim}{b}$, заполняет такое-то место отношения $R^i(,)$, которое является переводом слова $\underset{\sim}{a}$", будем говорить: "Слово $\underset{\sim}{b}$ заполняет такое-то место слова $\underset{\sim}{a}$".

---

[2] Отношение между дескрипторами $\{A_1, A_2, \ldots, A_n\}$, $n>1$, заполняющих одно и то же место отношения $R^i(,)$ - аналог отношения однородности [2].

Если два слова фразы $\underset{\sim}{a}$ и $\underset{\sim}{b}$ заполняют места одного и того же отношения $R^i(,)$, то будем говорить, что между словами $\underset{\sim}{a}$ и $\underset{\sim}{b}$ установлена связь. Если слово $\underset{\sim}{b}$ заполняет валентность слова $\underset{\sim}{a}$, то будем говорить, что между словами $\underset{\sim}{a}$ и $\underset{\sim}{b}$ установлена связь по валентности слова $\underset{\sim}{a}$. Связь по валентности называется сильной; все прочие связи называются слабыми, причем если слова $\underset{\sim}{a}$ и $\underset{\sim}{b}$ заполняют места предлога, то связь называется слабой предложной.

На входе алгоритм получает фразу, каждая словоформа которой снабжена морфологической информацией (в нашем эксперименте морфологическая информация приписывалась словоформе вручную) и каждое слово которой снабжено его словарной статьей (статьями) из ЕС-словаря. Кроме того, во фразе найдены так называемые предложные блоки, т.е. для каждого предлога установлено, каким существительным он управляет (поиск предложных блоков включен в эксперимент). Эта информация необходима потому, что валентности (места) некоторых слов могут быть заполнены (только) предложными блоками.

Алгоритм ЕС-анализа состоит из 5 циклов, каждый из которых устанавливает во фразе определенные связи. Первые три цикла устанавливают сильные связи, 4 цикл устанавливает слабые предложные связи, 5 цикл производит обработку однородных членов.

На каждом цикле фраза один раз просматривается слева направо[3].

Работа первых четырех циклов сводится по существу к тому, что для каждой пары слов РС (Рабочее Слово) и КЗ (Кандидат на Заполнение) проверяется, может ли КЗ заполнить какую-либо валентность (место) РС. Проверка организована следующим образом: сначала выясняется, удовлетворяет ли МИ КЗ'а хотя бы одному МТ какой-либо незаполненной валентности РС; при положительном ответе проверяется, удовлетворяет ли хотя бы один перевод КЗ хо-

[3] По техническим причинам операции "Предложного блока" и 1 цикла удобно выполнять за один пробег фразы.

тя бы одному СТ той же валентности (при этом учитывается, что МТ и СТ могут быть "привязаны" друг к другу). Если ответ положительный, то образуется формула.

Все циклы являются многовариантными, т.е. на всех циклах учитывается лексическая неоднозначность и неоднозначность связей (синтаксическая неоднозначность). Это означает, в частности, что при вышеописанной операции сравнения КЗ и РС каждая словарная статья КЗ сравнивается с каждой (незаполненной) валентностью РС (в каждой словарной статье РС). Благодаря этому более одной формулы выдается в таких, например, случаях: хорошее объединение - оценка (хороший, объединять), оценка (хороший, фирма); строит ВИНИТИ - субъект (ВИНИТИ, строить), объект (ВИНИТИ, строить) и т.п.

Если между КЗ и РС установлена связь и при этом записана более чем одна формула, то все эти формулы считаются конфликтными, т.е. такими, которые не могут быть в одной и той же СЗ. Подобно тому как строятся разные дорожки в синтаксическом анализаторе Куно - Эттингера [5], при появлении хотя бы одной пары конфликтных формул СЗ "расслаивается": образуется две или более СЗ, которые называются гипотетическими переводами (ГП) фразы. Одна из задач не рассматриваемого здесь этапа приведения СЗ - анализ и отсев ГП.

Заметим, что количество ГП одной фразы гораздо меньше, чем количество вариантов синтаксических структур, так как алгоритм ЕС-анализа использует не только синтаксические, но и семантические фильтры.

Если после работы всех пяти циклов у какого-нибудь слова остались незаполненные валентности, то они также переносятся из ЕС-словаря в СЗ фразы.

www.ruslang.ru

4.1. Итак, первые четыре цикла отличаются друг от друга только тем, какие слова фразы выбираются в качестве РС и КЗ. На 1-м цикле устанавливаются связи между существительным и согласованным с ним прилагательным. Прилагательное является РС, существительное - КЗ. Это диктуется тем, что прилагательное

67263

БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

"сильнее требует" существительное, чем существительное - прилагательное[4].

Итак, во фразе по очереди фиксируются прилагательные. Существительное - КЗ ищется справа от прилагательного, причем между прилагательным и существительным не должно быть ни одного слова, кроме прилагательных, наречий, сочинительных союзов, запятых (ср.: старый дом; старый, но очень удобный дом). Если такое существительное нашлось и оно может заполнить хотя бы одну валентность прилагательного, то образуется формула, во фразе фиксируется очередное прилагательное и т.д. Если же такого существительного нет или оно есть, но не может заполнить никакую валентность прилагательного, то существительное, удовлетворяющее этим же условиям, ищется слева от прилагательного (ср. ватин чистошерстяной, мешок спальный).

В описываемом варианте алгоритма предполагалось, что в анализируемых текстах нет окказиональных субстантивированных прилагательных.

На всех последующих циклах прилагательные словами не считаются (исключение составляют прилагательные, имеющие две или более валентностей, из которых одна может быть заполнена существительным, не согласованным с прилагательным, ср.: стакан, полный воды; такие прилагательные должны считаться словами на всех циклах).

4.2. На 2-м цикле в качестве РС и КЗ выбираются такие два слова, что 1) между ними нет ни одного слова a, которое не заполняет никакой валентности РС; 2) КЗ находится правее РС; 3) КЗ не связалось ни с одним словом фразы (кроме РС; напомним, что прилагательные не считаются словами и связь с ними не учитывается). На этом цикле КЗ может заполнить только незаполненную валентность РС. Предложный блок считается одним словом.

[4] Заметим, что прилагательные никогда не бывают КЗ, поэтому с точки зрения ЕС-анализа дескрипторам - переводам прилагательных не нужно приписывать семантические признаки. Тем не менее семантический признак приписывается каждому дескриптору - это необходимо для работы следующих этапов.

Итак, 2 цикл организован таким образом, что в СЗ, полученной в результате его работы, не может быть "пересечения" и "обрамления стрелок"[5].

Для большинства анализировавшихся фраз (напомним, что это заголовки и краткие резюме документов) первичная СЗ выдавалась после работы 2 цикла.

Примеры фраз, целиком проанализировавшихся после 2 цикла.

Проблема предохранения кранов от перегрузки

РС - проблема, КЗ - предохранения; РС и КЗ связались:

1. содержание (предохранять, проблема)

РС - предохранения, КЗ - кранов :

2. объект (кран, предохранять)

РС - то же, КЗ - от перегрузки:

3. причина (перегружать, предохранять)

Конструкция телескопической стрелы отечественных экскаваторов

На 1-м цикле получились формулы:

1. вид (телескопический, стрела)

2. источник (СССР, экскаватор)

Работа 2 цикла:

РС - конструкция, КЗ - стрелы :

3. параметр (конструкция, стрела)

РС - стрела, КЗ - экскаваторов :

4. часть (стрела, экскаватор)

Координационный план работы по решению проблемы 0. 80. 525

В этой фразе два омонима: работа (работать - "действие", публикация - "документ") и решение (решать - "действие", резолюция - "документ"). Слова 0. 80. 525 в ЕС-словаре нет. Словарная информация к нему создается особым вспомогательным алгоритмом, описание которого здесь не приводится; укажем только, что этот алгоритм использует информацию о морфологии цифровых комплексов (ср. 0. 80. 525 и I/2-72) и информацию

[5] Аналогом этого цикла в системе анализа по НС является установление некоторых непрерывных составляющих.

о переводе ближайшего к цифровому комплексу слова справа (ср. 1 апреля, 72 г., 30 автомобилей).

Фраза проанализировалась следующим образом:

ГП – 1

1. назначение (координировать, план)
2. содержание (работать, план)
3. содержание (решать, работать)
4. объект (проблема, решать)
5. индекс (0, 80, 525, проблема)

ГП – 2

1. – см. ГП – 1
2. содержание (публикация, план)
3. содержание (публикация, решать)
4. – см. ГП – 1
5. – см. ГП – 1

ГП – 3

1. – см. ГП – 1
2. содержание (публикация, план)
3. содержание (резолюция, публикация)
4. индекс (0, 80, 525, проблема) (см. ГП – 1)

Формула 2 во всех ГП получилась при: РС – план, КЗ – работы; формула 3 – при: РС – работы, КЗ – по решению; формула 4 в ГП – 1, 2 – при: РС – по решению, КЗ – проблемы; формула 5 – при: РС – проблемы, КЗ – 0, 80, 525.

4.3. Если после работы 1 и 2 циклов во фразе остались слова, не связавшиеся ни с одним другим, или если СЗ фразы не является связным графом, то фраза поступает на 3 цикл[6]. Во фразе по очереди фиксируются слова (кроме первого), которые не связались ни с одним из предыдущих слов. Каждое фиксированное слово является РС, а все слова, расположенные левее его, по очереди являются КЗ (предложный блок считается одним словом).

[6] Аналогом этого цикла в системе анализа по НС является установление некоторых непрерывных и некоторых прерывных составляющих.

Если хотя бы один КЗ может заполнить какую-либо валентность РС, не заполненную на предыдущих циклах, то образуется формула(формулы), а затем фиксируется очередное слово, не связавшееся ни с одним из предыдущих, и т.д. Если же ни одного такого КЗ не нашлось, то фиксированное слово становится КЗ, а слова, расположенные левее его, – РС. У РС проверяются валентности, не заполненные на предыдущих циклах. Предложный блок по-прежнему считается одним словом.

Как видно, в результате работы этого цикла может получиться "пересечение" и "обрамление стрелок". Кроме того, могут получиться такие ситуации: 1) одно и то же слово $\underset{\sim}{a}$ заполняет первое место валентностей нескольких слов $\underset{\sim}{b_1}, \underset{\sim}{b_2}, \ldots, \underset{\sim}{b_k}$, графически:



2) несколько слов $\underset{\sim}{b_1}, \underset{\sim}{b_2}, \ldots, \underset{\sim}{b_k}$ заполняют одну и ту же валентность слова $\underset{\sim}{a}$. В последних двух случаях полученные формулы являются конфликтными.

Нужно сказать, что МГ и СГ валентностей, как правило, весьма "жесткие", поэтому в результате работы 3 цикла количество ГП увеличивается относительно немного или даже совсем не увеличивается.

Примеры.

Просьба завода запланировать поставку в апреле УРИ-2

На 2 цикле были получены следующие формулы:

1. субъект (завод, просить)
2. содержание (планировать, просить)
3. объект (поставлять, планировать)

Ни у слова запланировать, ни у слова поставка нет валентности, которая могла бы заполниться предложным блоком в апреле, поэтому на 2 цикле он не связался ни с одним из слов, расположенных левее его; слово УРИ-2 не может заполнить ни одной валентности предложного блока в апреле (валентности предложного блока – это валентности существительного – члена предложного блока), поэтому на 2 цикле оно тоже не связалось ни с одним из слов, расположенных левее его. Следовательно, фраза поступает на 3 цикл. На

3 цикле: РС – в апреле; ни одно слово, расположенное левее его, не может заполнить его валентностей. КЗ – в апреле; не может связаться ни с одним словом, расположенным левее его, по валентности этого слова. РС – УРИ-2; валентностей у этого слова нет, следовательно: УРИ-2 – КЗ; связывается со словом поставка по его валентности:

4. объект (УРИ-2, поставлять).

Завод просит поставить УРИ-2

На 2 цикле были получены следующие формулы:

1. содержание (поставлять, просить)
2. объект (УРИ-2, поставлять)

На 3 цикле: РС – просит, КЗ – завод. Получаем формулу:

3. субъект (завод, просить).

О выделении магнитной ленты ВЦ Латвийского университета для БЭСМ-4, принадлежащей Госплану Латвии

Результаты работы 1 и 2 циклов:

ПП-1

1. вид (магнитный, лента)
2. локализация (Латвия, университет)
3. объект (лента, выделять)
4. субъект (ВЦ, выделять)
5. субъект (Госплан, иметь)
6. масштаб (Латвия, Госплан)

ПП-2

1. – см. ПП-1.
2. – см. ПП-1
3. – см. ПП-1
4. адресат (ВЦ, выделять)
5. – см. ПП-1
6. – см. ПП-1

Слова: для БЭСМ-4 и принадлежащей не связались ни с одним из предыдущих слов. Фраза поступает на 3 цикл. Результаты работы 3 цикла:

назначение (БЭСМ-4, выделять) (РС – выделение, КЗ – БЭСМ-4)
объект (БЭСМ-4, иметь)
объект (лента, иметь)
} формулы конфликтны (РС – принадлежащей, КЗ – БЭСМ-4, КЗ – лента).

Итак, получаем 4 ГП:

ГП-1

1. вид (магнитный, лента)
2. локализация (Латвия, университет)
3. объект (лента, выделять)
4. субъект (ВЦ, выделять)
5. субъект (Госплан, иметь)
6. масштаб (Латвия, Госплан)
7. назначение (БЭСМ-4, выделять)
8. объект (БЭСМ-4, иметь)

ГП-2

1. - 3. - см. ГП-1
4. адресат (ВЦ, выделять)
5. - 8. - см. ГП-1

ГП-3

1. - 7. - см. ГП-1
8. объект (лента, иметь)

ГП-4

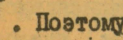
1. - 3. - см. ГП-1
4. адресат (ВЦ, выделять)
5. - 7. - см. ГП-1
8. объект (лента, иметь)

4.4. На 4 цикле обрабатываются несвязавшиеся предложные блоки. Предлог теперь считается отдельным словом. Напомним, что все предлоги переводятся отношениями. Задача 4 цикла - заполнить места предлога.

С технической точки зрения предлоги обрабатываются так же, как любые двувалентные слова. Предлог объявляется РС'ом; одно из его мест может быть заполнено только существительным - членом предложного блока; оно и объявляется КЗ. Строится формула $R^i(A,)$ (или $R^i(, A)$). Затем происходит поиск другого КЗ - слова, могущего заполнить другое место предлога. Поиск ведется слева от предлога, а при отрицательном ответе - справа: строится формула $R^i(, B)$ (или $R^i(B ,)$). На этапе приведения СЗ полученные формулы складываются по правилу

$$\frac{R^i(A, \ )\quad R^i(\ , B)}{R^i(A, B)}$$

Тот факт, что предпочтение отдается КЗ, находящемуся слева от предлога, объясняется тем, что анализировались пока только простые назывные предложения, ср. В Москве идут дожди - Осадки в Москве.

Как уже было сказано выше, в первичных СЗ некоторых (не всех) фраз со "слабыми предлогами" допустимы подграфы вида:  . Поэтому на этапе ЕС-анализа не считаются конфликтными формулы $R^i(A, B_1)$, $R^i(A, B_2)$, ..., $R^i(A, B_m)$, где B_1, B_2, ..., B_m - переводы разных слов фразы, если эти формулы получены на 4 цикле. Заметим, что подобная ситуация на 4 цикле получается очень часто. Это происходит потому, что место отношения, которое заполняется не членом предложного блока, обычно имеет "нежесткие" МТ и СТ; ср. МТ и СТ ко второму месту отношения локализация (,) - переводу предлога в : МТ - существительное или глагол, СТ - "действие" или "предмет".

Рассмотрим примеры.

Механизация строительства бетонных дорог в Западной Европе

Результат работы 1 и 2 циклов:

1. материал (бетон, дорога)
2. имя (Западный, Европа)
3. объект (строить, механизировать)
4. объект (дорога, строить)

Предложный блок в Европе обрабатывается на 4 цикле. Резуль-

тат работы 4 цикла:

5. локализация (Европа,)
6. локализация (, дорога)
7. локализация (, строить)
8. локализация (, механизировать)

Заметим, что ни одна из формул 6–8 не является лишней.

Борьба с голодом в Поволжье

На 2 цикле получена формула:

1. объект (голод, бороться)

На 4 цикле получены следующие формулы (для наглядности приведем сразу результат сложения):

2. локализация (Поволжье, голод)
3. локализация (Поволжье, бороться)

Заметим, что формула 3 здесь является лишней. Ср. совершенно аналогичную фразу:

Борьба с хулиганством в Москве

1. объект (хулиганство, бороться)
2. локализация (Москва, хулиганство)
3. локализация (Москва, бороться).

Здесь формула 3 не является лишней.

4.5. На 5 цикл поступает любая фраза, в которой есть сочинительный союз или запятая. Задача 5 цикла – найти во фразе однородные члены или установить, что запятая не является "однородной" (пока предполагается, что в анализируемых текстах нет и – частицы и, следовательно, не стоит задача отличать и – союз от и – частицы).

Организация 5 цикла существенно отличается от организации первых четырех циклов. А именно, сначала алгоритм вырабатывает гипотезы о рядах однородных членов[7], а затем эти гипотезы "пропускаются" через несколько "фильтров". В качестве "фильтров" используются, во-первых, некоторые общие свойства правильной синтаксической структуры и, во-вторых, специальные свойства предложения с однородными членами.

---

[7] Рядом однородных членов называются слова $a_1, a_2, \ldots, a_n$ такие, что их переводы $A_1, A_2, \ldots, A_n$ заполняют одно и то же место отношения $R^i(,)$: $R^i(\{A_1, A_2, \ldots, A_n\}, B)$.

Рассмотрим первую группу "фильтров".

1) Для синтаксической структуры фразы справедливо, что слова, управляющие или управляемые рядом однородных членов, находятся в предложении по одну сторону ряда однородных. Аналогичное утверждение справедливо и для первичной СЗ.

2) Для синтаксической структуры фразы установлено, что если слова $a_1, a_2, \ldots, a_p$ являются рядом однородных членов, а слова $b_1, b_2, \ldots, b_q$ являются другим рядом однородных членов, то невозможно такое расположение их во фразе: $a_m\ b_i\ a_n\ b_j$. Аналогичное утверждение справедливо и для первичной СЗ.

3) Если в системе синтаксического анализа используется понятие синтаксической валентности (см., например, [1]), то в ситуации, когда одному и тому же предложению сопоставляется более одной структуры, предпочтение отдается структуре с насыщенными валентностями. Это утверждение (или его аналог для СЗ) является справедливым, если рассматриваются валентности слов, которые сами не являются членами ряда однородных. Данное свойство используется как "фильтр" при рассмотрении слов - не однородных членов. О заполненности - незаполненности валентностей однородных членов см. ниже.

Сформулируем специальные свойства первичной СЗ фразы с однородными членами (вторая группа "фильтров"). Эти свойства, так же как и свойства правильной синтаксической структуры [1], были получены эмпирически. Для каждого свойства будут приведены две фразы с таким минимальным различием, что если одной из них сопоставить правильную СЗ, то для другой эта же СЗ окажется неправильной. Так как выбранные фразы отличаются только одним признаком, правильность СЗ в одном случае и неправильность ее в другом можно объяснить только влиянием этого признака. Поэтому естественно считать, что правильная СЗ должна обладать свойством, учитывающим этот признак[8].

4) Если фразе может быть сопоставлено несколько ПП, различающихся по признаку "заполненность - незаполненность валент-

---

[8] Этот абзац является почти дословным пересказом одного из абзацев [1], стр. 11.

ностей членов ряда однородных", то правильным считается тот ГП, в котором либо все члены ряда однородных имеют заполненные валентности, либо все - не имеют.

Пример.

Никогда не видел сырой земли и болот

Никогда не видел сырой земли лесов и болот

В первой фразе однородные члены - земли и болот. Во второй фразе у слова земли есть заполненная валентность: принадлежность(земля, лес), а у слова болот заполненных валентностей нет. По этой причине слова земли и болот во второй фразе не являются однородными членами.

5) Если фразе может быть сопоставлено несколько ГП и в некоторых из них переводы слов - членов ряда однородных имеют разные семантические признаки, то правильным считается тот ГП, в котором переводы однородных членов имеют одинаковые семантические признаки (два семантических признака "a" и "b" такие, что "a"$\subset$"b", считаются одинаковыми).

Рассмотрим это свойство на примере фразы Наука и техника за рубежом. Слово техника имеет три перевода. Один из них имеет семантический признак "устройство" (ср. На этом заводе постоянно внедряют новую технику); другой перевод имеет семантический признак "параметр (,)" (ср. За последние 100 лет техника кондиционирования воздуха изменилась кардинальным образом); третий перевод имеет семантический признак "область деятельности" (ср. Развитие техники требует хороших инженеров). Слово наука имеет один перевод, семантический признак которого - "область деятельности". Итак, рассматриваемому словосочетанию может быть сопоставлено три ГП. Выбирается ГП с тем переводом слова техника, который имеет семантический признак "область деятельности".

Оба сформулированных свойства фраз с однородными членами являются независимыми друг от друга, т.е. для каждого свойства можно найти неправильную СЗ, обладающую всеми свойствами правильной СЗ, кроме рассматриваемого - см. примеры выше.

Рассмотрим, как формируются гипотезы о рядах однородных членов. Фраза делится на непересекающиеся и не вкладывающиеся друг

в друга сегменты, границами которых служат начало фразы, каждый сочинительный союз, каждая запятая, конец фразы. В каждом сегменте фиксируется слово $\underset{\sim}{a}$, которое не заполняет первое место ни в одном отношении $R^i(,)$ (т.е. такое слово, в которое "не входит стрелка").

Как уже было сказано выше, в СЗ фразы с однородными членами должна быть хотя бы одна формула вида $R^i(\{A_1, A_2, \dots, A_n\}, \{B_1, B_2, \dots, B_m\})$, где $n>1$ или $m>1$. До работы 5 цикла в первичной СЗ фразы таких формул нет, а есть формулы вида $R^i(A_j, B_k)$. В результате работы 5 цикла к дескриптору $A_j(B_k)$ приписываются другие дескрипторы, так что место отношения оказывается заполненным более чем одноэлементным множеством дескрипторов.

Отсюда понятно, что для формирования гипотезы необходима следующая проверка: может ли фиксированное слово $\underset{\sim}{a}$ заполнить какое-нибудь место отношения $R^i(,)$ такое, что оно - в результате работы предыдущих циклов - уже заполнено словом $\underset{\sim}{b}$. При положительном ответе формируется гипотеза о том, что слова $\underset{\sim}{a}$ и $\underset{\sim}{b}$ являются членами одного и того же ряда однородных.

Для анализа резюме и заголовков, по-видимому, наиболее целесообразно включить фильтр - свойство (I) в формирование гипотез. Для разбираемых ниже примеров это так и сделано.

Примеры (знак "/" обозначает границу сегмента; стрелки между словами обозначают отношения, которыми связались слова на 1-4 циклах):



Гипотеза 1

I. отбор и уничтожение: содержание ({отбирать, уничтожать}, делать)

Гипотеза 2

I. на хранение и уничтожение: цель ({хранить, уничтожать}; отбирать)

Выбирается Гипотеза 2 - см. фильтр 4.

План разработки системы автоматического сбора / и передачи первичной информации

Гипотеза 1

1. разработки и передачи: содержание ({ разрабатывать, передавать }, план)

Гипотеза 2

1. сбора и передачи: содержание ({ собирать, передавать }, система), объект (информация, { собирать, передавать })[9]

Выбирается Гипотеза 2 – см. фильтр 3.

План по механизации учетно-статистических работ / и внедрения в народное хозяйство вычислительной техники

Гипотеза 1

1. по механизации и внедрения: содержание ({ механизировать, внедрять }, план) (валентность содержание (, план) имеет два МТ: ср., план по механизации, план внедрения.) Эта гипотеза правильно отражает смысл фразы.

Гипотеза 2

1. работ и внедрения: объект ({ работать, внедрять }, механизировать) (Примеч.: валентность объект (, механизировать) имеет СТ "действие").

Обе гипотезы "проходят через все фильтры".

Доклад о производстве насосного оборудования для механизации водоподъема / и опреснения воды / и об улучшении эксплуатации водопойных сооружений / и источников водоснабжения на пастбищах

Гипотеза 1

1. для механизации и опреснения
2. о производстве и об улучшении
3. сооружений и источников

Гипотеза 2

1. – см. Гипотезу 1
2. – см. Гипотезу 1
3. эксплуатации и источников

[9] Если у одного из "кандидатов на однородность" есть заполненная валентность, а у других "кандидатов" одноименная валентность не заполнена, то такая валентность заполняется тем же словом (в соответствии с фильтром 1).

Гипотеза 3

1. - см. Гипотезу 1
2. - см. Гипотезу 2
3. воды и источников

Гипотеза 4

1. водоподъема и опреснения
2. - см. Гипотезу 1
3. - см. Гипотезу 2

Гипотеза 5

1. водоподъема и опреснения
2. - см. Гипотезу 1
3. воды и источников

Гипотеза 6

1. - см. Гипотезу 4
2. - см. Гипотезу 1
3. - см. Гипотезу 1

"Фильтр" 2 отбрасывает Гипотезу 3. "Фильтры" 3 и 4 не отбрасывают ни одной гипотезы. "Фильтр" 5 отбрасывает Гипотезы 2, 4 (см. ряд 3) и Гипотезу 5 (см. ряд 3). Оставшиеся Гипотезы 1 и 6 удовлетворяют всем требованиям.

4.6. Результаты машинного эксперимента по ЕС-анализу. В 1972 году алгоритм ЕС-анализа был запрограммирован (1-4 блоки) в ВЦ НИИСтроммаш (г. Гатчина) для ЭЦВМ "Минск-22". В настоящее время алгоритм отлаживается на экспериментальном массиве заголовков. Одновременно программируется 5 цикл и цикл разрешения неоднозначности СЗ фразы.

Ниже приводятся некоторые полученные первичные СЗ. Для большинства проанализированных фраз окончательная СЗ совпадает с первичной СЗ из-за простоты исходного материала (привлекались к анализу пока только заголовки, представляющие собой именные конструкции без однородных членов и обособленных оборотов). Усложнение структуры анализируемых фраз, которое неизбежно при переходе к анализу текста реферата (а следовательно, любого произвольного текста), потребует детальной разработки этапа собственно семантического анализа (СС-анализа). Задачей этого этапа является преобразование первичной СЗ в некоторую нормализованную окончательную СЗ текста, которая, в свою очередь, будет преобразована в информационную запись (ИЗ) текста.

5. Заключение.

Итак, построение ИЗ текста предполагается осуществлять в три этапа:

I этап. Переход от текста к первичной СЗ этого текста.

II этап. Преобразование первичной СЗ в приведенную СЗ текста.

III этап. Преобразование приведенной СЗ текста в ИЗ текста.

В данной статье описан способ выполнения задачи I этапа (пока в неполном объеме). На этом этапе (этапе ЕС-анализа) происходит перевод каждой фразы с ЕЯ на ИЯ на основе информации, зафиксированной в ЕС-словаре, и некоторых общих свойств правильной синтаксической структуры фразы. ИЯ ЭХО устроен таким образом, что этот перевод может происходить практически без потерь информации (некоторые виды информации, например, о логическом акценте фразы или о грамматическом числе существительных, сознательно утрачиваются как избыточные для построения ИЗ текста). Запись на ИЯ, получаемая в результате ЕС-анализа (первичная СЗ), не является еще правильной структурой: в ней могут быть разрывы, непереведенные слова, альтернативные формулы. Основная задача ЕС-анализа - не столько перевести на ИЯ слова-дескрипторы, сколько разместить их на соответствующих местах смысловых отношений. Смысловые отношения являются переводами на ИЯ единиц разных лингвистических уровней: внутрисловных связей, связей сильного управления, слабого управления, некоторых грамматических категорий, а также некоторых слов ЕЯ (не только служебных элементов, но и некоторых полнозначных слов, например, причина, больше, равно, сходно и др.). При переходе к анализу текста межфразовые связи тоже будут передаваться в СЗ смысловыми отношениями. Такое выравнивание структуры текста обеспечивает очень ценное для наших задач свойство СЗ - ее эластичность. В первую очередь именно это свойство позволяет нам выполнять разные формальные процедуры над смыслом целого текста, и не только над смыслом, выраженным в отдельном предложении. Одна из наиболее благодарных задач, которые можно решать с помощью СЗ такого вида - автоматическое сжатие информации, содержащейся в исходном документе, для записи ее в машине

и последующего поиска. Сохранение в СЗ всей фактической информации исходного текста и возможность вычленять из СЗ любой ее фрагмент позволяет осуществлять по любому заданному тексту узкофактографический поиск, а по комплексу текстов – систематизацию информации, смысловые выводы и синтез необходимой части информации.

Приложение. Первичные СВ фраз, полученные на машине.

ОПРЕДЕЛЕНИЕ ЕМКОСТИ АВТОМАТИЧЕСКИХ ГРЕЙФЕРОВ

1АВТОМАТИЧЕСКИЙ3 ВИД3 ГРЕЙФЕР4

2ЕМКОСТЬ2 ОБЪЕКТ1 ОПРЕДЕЛЯТЬ1

2ЕМКОСТЬ2 ПАРАМЕТР2 ГРЕЙФЕР4

АВТОМАТИЗАЦИЯ РАБОТЫ ГИДРАВЛИЧЕСКИХ ПРИВОДОВ ЭКСКАВАТОРА

1ГИДРАВЛИЧЕСКИЙ3 ПРИНЦИП2ДЕЙСТВ ПРИВОД4

2РАБОТАТЬ2 ОБЪЕКТ1 АВТОМАТИЗИРОВАТЬ1

2ПРИВОД4 СУБЪЕКТ4 РАБОТАТЬ2

2ПРИВОД4 ЧАСТЬ6 ЭКСКАВАТОР5

ПРОБЛЕМЫ ПРЕДОХРАНЕНИЯ КРАНОВ ОТ ПЕРЕГРУЗКИ

2ОХРАНЯТЬ2 СОДЕРЖАНИЕ1 ПРОБЛЕМА1

2КРАН3 ОБЪЕКТ2 ОХРАНЯТЬ2

2ПЕРЕГРУЗКА6 ПРИЧИНА3 ОХРАНЯТЬ2

КОНСТРУКЦИЯ ТЕЛЕСКОПИЧЕСКОЙ СТРЕЛЫ ОТЕЧЕСТВЕННЫХ ЭКСКАВАТОРОВ

1ТЕЛЕСКОПИЧЕСКИЙ2 ВИД2 СТРЕЛА3

1СССР4 ИСТОЧНИК5 ЭКСКАВАТОР5

2КОНСТРУКЦИЯ1 ПАРАМЕТР1 СТРЕЛА3

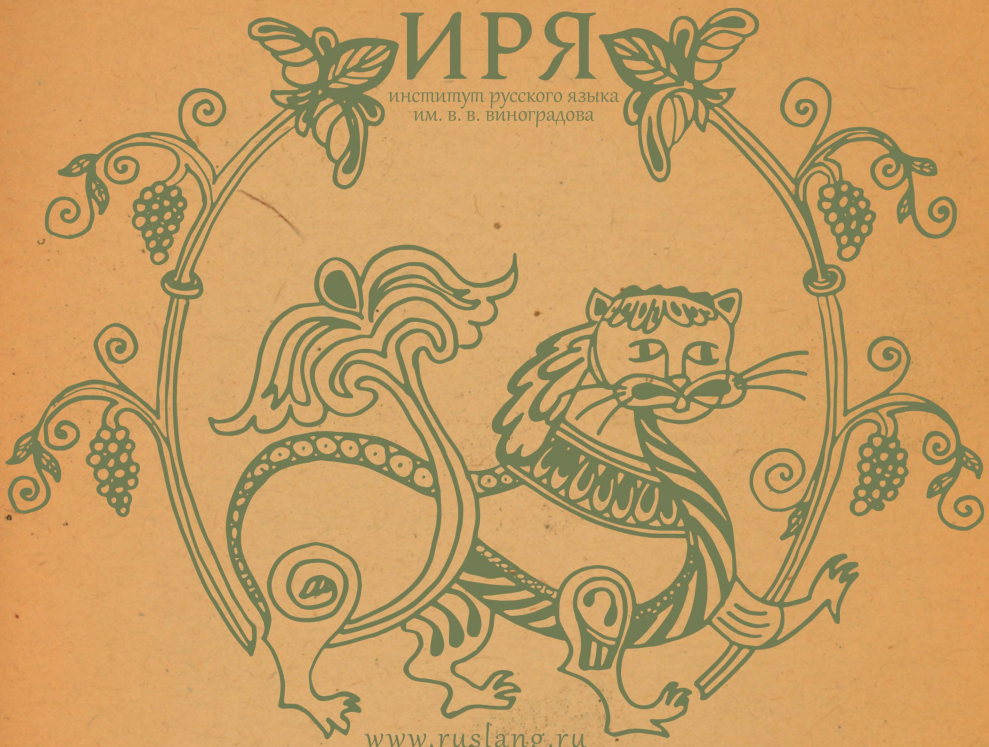
2СТРЕЛА3 ЧАСТЬ3 ЭКСКАВАТОР5

Литература

1. Л.Н. Иорданская. Автоматический синтаксический анализ. т.II. Межсегментный синтаксический анализ. Новосибирск, 1967.
2. Ю.А. Шрейдер. О свойствах проективности языка. - "Научно-техническая информация", 1964, № 8.
3. Н.Н. Леонтьева. Создание информационного языка на базе семантического анализа текста. - "Научно-техническая информация", сер. 2, 1971, № 8.
4. А.Е. Гуревич. Автоматическое определение неконфликтности результатов семантического анализа русских предложных групп. - "Сборник научных работ № 19. Прикладная лингвистика." Мос. пед. ин-т иностр. яз., Студенческое научное общество. М., 1972.
5. С. Куно, А. Этингер. Многовариантный синтаксический анализатор. Пер. с англ. - "Автоматический перевод". М., 1971.
6. З.М. Шаляпина. Семантические элементы как основа лексикографического описания: общие принципы и формальный аппарат.- "Машинный перевод и прикладная лингвистика", № 15, М., 1973.

БИБЛИОТЕКА
ИН-ТА РУССКОГО ЯЗЫКА
ИН-ТА ЯЗЫКОЗНАНИЯ АН СССР

ИРЯ
институт русского языка
им. в. в. виноградова



www.ruslang.ru

А04048 от 16.II.1973 г. 150 экз. 2,25 п.л. Зак. 439

Тип МГПИИЯ