МЕЖДУНАРОДНЫЙ СТАНДАРТ

ISO 24615-1

Первое издание
2014-02-01

Управление языковыми ресурсами. Система синтаксического аннотирования (SynAF).

Часть 1. **Синтаксическая модель**

Language resource management. — Syntactic annotation framework (SynAF) —
Part 1: Syntactic·model

Ответственность за подготовку русской версии несёт GOST R (Российская Федерация) в соответствии со статьёй 18.1 Устава ISO

Ссылочный номер
ISO 24615-1:2014

Содержание

Страница

Предисловие

Международная организация по стандартизации (ISO) является всемирной федерацией национальных организаций по стандартизации (комитетов-членов ISO). Разработка международных стандартов обычно осуществляется техническими комитетами ISO. Каждый комитет-член, заинтересованный в деятельности, для которой был создан технический комитет, имеет право быть представленным в этом комитете. Международные правительственные и неправительственные организации, имеющие связь с ISO, также принимают участие в работе. ISO работает в тесном сотрудничестве с Международной электротехнической комиссией (IEC) по всем вопросам стандартизации в области электротехники.

Процедуры разработки, использованные при создании настоящего документа, и надлежащие операции его дальнейшего сопровождения описаны в Части 1 Директив ISO/IEC, в которых особенно важны критерии утверждения и приёмки различных документов ISO. Проект настоящего документа был подготовлен в полном соответствии с правилами, приведёнными в Части 2 Директив ISO/IEC (см. www.iso.org/directives).

Принимается во внимание тот факт, что некоторые из элементов данного документа могут быть объектом патентных прав, однако ISO не принимает на себя никаких обязательств по определению отдельных или всех таких прав. Подробные сведения о патентных правах, выявляемых в процессе разработки того или иного документа, подлежат включению в его введение и/или в перечень полученных ISO деклараций (см. www.iso.org/patents).

Любая торговая марка, фигурирующая в настоящем документе, является информацией, представляемой исключительно для удобства пользователей стандарта, и не может считаться признаком выражаемого предпочтения.

Разъяснение специфических терминов и формулировок ISO, касающихся оценки соответствия конкретным нормативным документам, равно как и приверженности ISO принципам Всемирной торговой организации (ВТО) в части устранения технических барьеров в торговле, можно найти на сайте ISO по унифицированному указателю информационного ресурса (URL) Foreword — Supplementary information.

Настоящий документ был подготовлен подкомитетом SC 4 “*Управление языковыми ресурсами*” Технического комитета ISO/TC 37 “*Терминология и другие языковые и информационные ресурсы*”.

Настоящее первое издание ISO 24615-1 отменяет и заменяет собой стандарт ISO 24615:2010, претерпевший минимальные изменения.

ISO 24615 (во всех его частях) предназначен для использования в сочетании со стандартами ISO 24612 “*Управление языковыми ресурсами. Система лингвистического аннотирования (LAF)*”, ISO 24613:2008, “*Управление языковыми ресурсами. Схема лексической разметки (LMF)*” и ISO 24611 “*Управление языковыми ресурсами. Система морфосинтаксического аннотирования (MAF)*”.

ISO 24615 состоит из следующих частей, объединённых общим заголовком *“Управление языковыми ресурсами. Система синтаксического аннотирования (SynAF)”*:

— *Часть 1. Синтаксическая модель*

и находящаяся в стадии разработки

— *Часть 2. Индексирование XML-документов (<Tiger2/>).*

Введение

В основе стандартов серии ISO 24615 лежат многочисленные проекты и мероприятия последнего десятка лет, предшествовавшие стандартизации [9] и обеспечившие создание различных моделей и форматов представления синтаксической информации в виде механизмов синтаксического анализа или аннотирования языковых ресурсов (корпусов с синтаксической разметкой). В течение целого ряда лет стандартом де-факто для процедур синтаксического аннотирования служил инициативный проект Penn Treebank (синтаксически аннотированный текстовый корпус Пенсильванского университета). Однако в более поздних работах — таких как инициативный проект Negra/Tiger в Германии (см. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/) и проект ISST в Италии [18] — были продемонстрированы более связные, более согласованные и более понятные системы синтаксического анализа, способные в равной степени учитывать как отношения иерархической соподчинённости компонентов, так и явление зависимости в синтаксическом аннотировании.

Особенно плодотворным в этом смысле оказался проект электронных информационных ресурсов “LIRICS”, который позволил объединить усилия группы экспертов, инициировавшей разработку проекта Международного стандарта ISO 24615, касающегося системы синтаксического аннотирования (SynAF). Эта группа в процессе разработки системы SynAF установила тот факт, что в действительности в рамках всех перечисленных выше инициативных проектов используется общая модель данных, которая создаёт хорошую основу для построения метамодели системы синтаксического аннотирования [см. результаты исследований, представленные в информационном бюллетене Deliverable D.3.1 по проекту Европейского Союза LIRICS; эта публикация касается оценки инициативных проектов систем морфосинтаксического и синтаксического аннотирования (“Evaluation of initiatives for morpho-syntactic and syntactic annotation”) и размещена по адресу http://lirics.loria.fr/doc_pub/Del3_1_V2.pdf].

В настоящей части ISO 24615 предлагается метамодель синтаксического аннотирования совместно с перечнем релевантных категорий данных для синтаксического аннотирования. Соответствующие категории данных доступны на сервере ISOCat (http://www.isocat.org/) в синтаксическом профиле (как он определён в ISO 12620:2009).

Управление языковыми ресурсами. Система синтаксического аннотирования (SynAF).

Часть 1.
Синтаксическая модель

1 Область применения

Первая часть ISO 24615 содержит описание системы синтаксического аннотирования (SynAF), которая является высокоуровневой моделью для представления синтаксической аннотации лингвистической информации в целях обеспечения надлежащего взаимодействия между языковыми ресурсами или компонентами процесса обработки языковых данных. Эта часть ISO 24615 служит дополнением стандарта ISO 24611 (относящегося к системе морфосинтаксического аннотирования) и тесно связана с ним, поскольку предоставляет метамодель для формирования синтаксических представлений и эталонные категории данных для отображения местоположения и зависимости информационных объектов в рамках сложных предложений или других сопоставимых высказываний и сегментов.

2 Нормативные ссылки

Перечисленные ниже ссылочные документы обязательны для применения данного документа. В случае датированных ссылок действующим является только указанное издание. Применительно к недатированным ссылочным документам применяются их самые последние издания (включая все последующие изменения):

ISO 1087-1:2000, *Терминологическая работа. Словарь. Часть 1. Теория и применение*

ISO 12620:2009, *Терминология, другие языковые ресурсы и ресурсы содержания. Спецификация категорий данных и ведение реестра категорий данных для языковых ресурсов*

ISO 24611:2012, *Управление языковыми ресурсами. Система морфосинтаксического аннотирования*

3 Термины и определения

Для целей данного документа используются термины и определения из стандартов ISO 1087-1:2000, ISO 12620:2009, ISO 24611:2012, а также терминология, приведённая ниже.

3.1
обстоятельственное слово, обстоятельство, адъюнкт
adjunct
второстепенный элемент, ассоциируемый с глаголом, в отличие от *синтаксических аргументов* (3.19)

Примечание 1 к словарной статье: Обстоятельственными словами в предложении могут быть наречия.

3.2
фрагмент
chunk
нерекурсивная *составляющая* (3.4)

3.3
клауза, элементарное предложение
clause
группа *фраз* (3.14), обычно содержащая некоторое высказывание

Примечание 1 к словарной статье: Клауза может быть *главным предложением* (3.10) или *подчинённым предложением* (3.17). В тех языках, где существует понятие законченности действия, клаузы с глагольным сказуемым могут быть как законченными, так и незаконченными — в зависимости от конкретной формы глагола. Главное предложение может само по себе являться законченным *предложением* (3.15). В модели SynAF клауза представляет собой особый случай *конституенты* (3.4).

3.4
конституента, составляющая
constituent
синтаксическая группировка слов [во *фразе* (3.14)], фраз [*в предложении* (3.3) либо в *другой фразе*] или элементарных предложений [в сложном *предложении* (3.15)] на основе выделения их структурных (или иерархических) свойств

3.5
зависимость, отношение зависимости
dependency
dependency relation
синтаксическая связь между *словоформами* (3.24) или *конституентами* (3.4), устанавливаемая на основе *грамматических функций* (3.7), которые выполняются *конституентами* относительно друг друга

3.6
(синтаксическая) дуга
syntactic edge
edge
тройка элементов, образуемая исходным *узлом* (3.12), целевым *узлом* и не обязательными *аннотациями* (3.9)

Примечание 1 к словарной статье: *Нетерминальные узлы* (3.13) имеют исходящую дугу синтаксической конституентности.

3.7
грамматическая функция
grammatical function
грамматическая роль, выполняемая *словоформой* (3.24) или *конституентой* (*3*.4) в рамках своего синтаксического окружения

Примечание 1 к словарной статье: Например, именная группа (NP) или имя существительное может играть внутри *предложения* (3.15) роль подлежащего — соответственно положению глагола в схеме графа отношения подчинённости. Между именной группой как подлежащим и основным глаголом предложения существует грамматическая связь. Все грамматические отношения (подлежащее — сказуемое, главное слово — модификатор и др.) категоризируются в соответствии с концепцией *отношения зависимости* (3.5) между терминальными или нетерминальными узлами.

3.8
(синтаксическая) вершина, главное слово
syntactic head
head
та часть *конституенты* (3.4), которая определяет область её действия (конкретное синтаксическое окружение, в котором может появляться конституента) и грамматические характеристики (например, если главное слово — женского рода, то грамматический род конституенты в целом тоже будет женским)

Примечание 1 к словарной статье: Присутствие главного слова конституенты обязательно.

3.9
(лингвистическая) аннотация
linguistic annotation
annotation
пара “характеристика — значение”, представляющая лингвистическое свойство лингвистического сегмента

3.10
главное предложение
main clause
клауза (3.3), которая может сама по себе быть законченным *предложением* (3.15)

Примечание 1 к словарной статье: В языках, предусматривающих различение завершённости и незавершённости действия, главное предложение обычно бывает законченным высказыванием. Пример: *Поезд опаздывает.*

3.11
модификатор
modifier
часть *конституенты* (3.4), описывающая свойство её *главного слова* (3.8)

Примечание 1 к словарной статье: Модификатор может располагаться до или после главного слова (вершины) *фразы* (3.14), являясь пре-модификатором или пост-модификатором, соответственно. Модификаторы в конституенте не обязательны.

3.12
(синтаксический) узел
node
syntactic node
словоформа (3.24) или *конституента* (3.4), рассматриваемая как элементарный синтаксический компонент дерева синтаксического анализа

3.13
нетерминальный узел
non-terminal node
синтаксический узел (3.12), не являющийся *словоформой* (3.24)

Примечание 1 к словарной статье: Нетерминальный узел имеет исходящую дугу коституентности (3.6).

3.14
фраза, синтаксическая конструкция
phrase
группа *словоформ* (3.24) (обычно состоящая из одного или нескольких слов), которая может выполнять определённую *грамматическую функцию* (3.7), например, в элементарном *предложении* (3.3)

Примечание 1 к словарной статье: Допускается присутствие пустых фраз (представленных неопределённо-личными местоимениями, которые в английском языке иногда снабжаются пометой “pro” и в элементарных предложениях выступают в роли подлежащего). Фраза, как правило, именуется по её *главному слову* (3.8): например, могут быть именные группы, глагольные группы, группы прилагательного, наречные группы и предложные группы. В просторечии фразы характеризуются как “раздутые слова” — в том смысле, что части фразы, добавляемые к главному слову, усложняют и конкретизируют его референцию. В нашей модели фраза является особым случаем *конституенты* (3.4).

3.15
сложное предложение
sentence
связанная группа *словоформ* (3.24), содержащая предикацию, которая обычно выражает законченную мысль и образует базовую единицу структуры дискурса

Примечание 1 к словарной статье: Сложное предложение состоит из одной или нескольких *клауз* (3.3). При описании речевого общения обычно говорят не о предложениях, а о “высказываниях”.

3.16
интервал
span
пара точек [p1, p2] (где $p1 \leqslant p2$), идентифицирующая сегмент документа, к которому применима *аннотация* (3.9)

Примечание 1 к словарной статье: Кратный интервал — это последовательность интервалов, в которой координаты конечной точки каждого предшествующего интервала меньше или равны координатам начальной точки последующего интервала.

3.17
придаточное предложение
subordinate clause
элементарное предложение, выполняющее некоторую *грамматическую функцию* (3.7) в *синтаксической конструкции* (3.14) [например, функцию определительного *предложения* (3.3), которое модифицирует существительное именной группы, образующее синтаксическую *вершину* (3.8)] или в другом элементарном предложении

Примечание 1 к словарной статье: Придаточное предложение обычно не выражает законченную мысль, а является частью более длинного сложного предложения**.**

3.18
фрейм субкатегоризации
subcategorization frame
совокупность ограничений, показывающих свойства *синтаксических аргументов* (3.19), которые могут или должны связываться с глаголом

ПРИМЕР Альфред (/syntacticArgument*1*) читает книгу (/syntacticArgument*2*) сегодня (/adjunct*1*).

Примечание 1 к словарной статье: Подлежащее, косвенное дополнение и прямое дополнение представляют собой субкатегоризированные *грамматические функции* (3.7) внутри предложения; все они подчиняются глаголу (т.е. могут появляться во фреймах субкатегоризации).

3.19
синтаксический аргумент
syntactic argument
важный функциональный элемент, запрашиваемый и интерпретируемый вершиной его *синтаксической конструкции* (3.14) или *узлом* (3.12), от которого он зависит (примером может служить именной аргумент предложной группы или глагол)

Примечание 1 к словарной статье: Применительно к глаголам и глагольным конструкциям аргументы идентифицируют стороны процесса, указываемого глаголом. В некоторых системах синтаксические аргументы называются дополнениями.

3.20
(синтаксический) граф
syntactic graph
graph
связное множество *синтаксических узлов* (3.12) и *дуг* (3.6)

3.21
синтаксическое дерево
syntactic tree
синтаксический граф (3.20), в котором каждый из узлов имеет единственный родительский узел

3.22
синтаксис, синтаксические правила
syntax
способ соединения и/или группирования *словоформ* (3.24) в грамматические обороты для сбора информации о существующих отношениях между группируемыми единицами

3.23
терминальный узел
terminal node
синтаксический узел (3.12), являющийся одиночной *словоформой* (3.24) или пустым элементом синтаксического отношения

3.24
словоформа
word form
непрерывный или сегментированный объект речевого или текстового оборота, идентифицируемый как автономная лексема.

4 Метамодель SynAF

4.1 Вводные замечания

В обработке языковых данных синтаксические аннотации выполняют как минимум две функции:

a) представление лингвистической конституентности подобно именным группам (NP), описывающим структурированную последовательность морфосинтаксически аннотированных лексем (включая пустые элементы или следы, порождённые перемещениями на уровне конституент), а также построение конституент из сегментированных элементов;

b) представление отношений зависимости, таких как “главное слово — модификатор” и отношения между категориями одного вида (подобные связям между главными словами в именных аппозициях или именным соподчинениям в некоторых формализмах). Внутри синтаксической группы может существовать информация о зависимости между элементами, прошедшими этап морфосинтаксического аннотирования (например, прилагательное может быть модификатором главного существительного внутри именной группы) или может описываться конкретное отношение между синтаксическими составляющими на клаузальном и пропозициональном уровнях (т.е. там, где именная группа выступает как “субъект” основного глагола элементарного или сложного предложения). Отношение зависимости может устанавливаться также для пустых элементов (например, для элемента “pro” в романских языках, где он выполняет грамматическую функцию).

Как следствие, синтаксические аннотации должны соответствовать многоуровневой стратегии, обеспечивающей взаимосвязь синтаксического аннотирования по составляющим элементам и по отношениям зависимости, как это установлено в метамодели SynAF.

4.2 Описание метамодели SynAF

4.2.1 Общий обзор

Метамодель SynAF представляется как совокупность классов универсального языка моделирования UML, дополненная UML-парами “атрибут — значение”, которые представляют соответствующие категории синтаксических данных. Текстовые описания метамодели SynAF определяют более полную информацию о её классах, отношениях и расширениях, которые могут быть включены в диаграмму UML. Выбор категории данных (DCS) осуществляется разработчиками в соответствии с установленными для SynAF процедурами (см. Рисунок 1). Для представления синтаксических аннотаций должны использоваться категорий данных, указанные в Приложении А.

4.2.2 Класс SyntacticNode

SyntacticNode — это параметризованный класс, категоризирующий как класс терминальных узлов, так и класс нетерминальных узлов. Синтаксические узлы могут быть задействованы в любом необходимом числе синтаксических отношений (см. 3.6 **синтаксические дуги**).

4.2.3 Класс T_Node

Класс *T_Node* представляет терминальные узлы синтаксического дерева, состоящего из словоформ, прошедших этап морфосинтаксического аннотирования, и из пустых элементов, когда они необходимы. Узлы этого класса определяются на одном *интервале* или на множестве интервалов (множественные интервалы обеспечивают учёт нарушений непрерывности составляющих частей текста). Для аннотирования узлов класса T_Node используются средства автоматической синтаксической категоризации, действующие на уровне отдельных слов.

4.2.4 Класс NT_Node

Класс *NT_Node* представляет нетерминальные узлы синтаксического дерева. Синтаксические деревья состоят в основном из узлов классов T_Node и NT_Node, а также из пустых элементов, когда они необходимы. Узлы класса T_Node реализуют обращения к интервалам, благодаря чему с помощью древовидного синтаксического представления могут быть получены интервалы и для узлов класса NT_Node. Для аннотирования узлов класса NT_Node используются средства автоматической синтаксической категоризации, действующие на уровне фраз и на более высоких уровнях (клаузальном и сентенциальном).

4.2.5 Класс SyntacticEdge

Класс *SynacticEdge* представляет отношение между синтаксическими узлами (как терминальными, так и нетерминальными). Например, отношение зависимости — это бинарное отношение, образуемое парой узлов — исходным и целевым — с одной или несколькими аннотациями. В частности,