МЕЖДУНАРОДНЫЙ СТАНДАРТ

ISO 24612

Первое издание
2012-06-15

Управление языковыми ресурсами. Лингвистическая аннотационная система (LAF)

Language resource management. – Linguistic annotation framework (LAF)

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

Ссылочный номер
ISO 24612:2012(R)

ДОКУМЕНТ ЗАЩИЩЁН АВТОРСКИМ ПРАВОМ

Бюро ISO по авторским правам:
Case postale 56 • CH-1211 Geneva 20
Тел.: + 41 22 749 01 11
Факс: + 41 22 749 09 47
Эл. почта: copyright@iso.org
Веб-сайт: www.iso.org

Опубликовано в Швейцарии

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.

Проекты международных стандартов разрабатываются согласно правилам, приведённым в Директивах ISO/IEC, Часть 2.

Разработка международных стандартов является основной задачей технических комитетов. Проекты международных стандартов, принятые техническими комитетами, рассылаются комитетам-членам на голосование. Для публикации в качестве международного стандарта требуется одобрение не менее 75 % комитетов-членов, принявших участие в голосовании.

Принимается во внимание тот факт, что некоторые из элементов настоящего документа могут быть объектом патентных прав. ISO не принимает на себя обязательств по определению отдельных или всех таких патентных прав.

ISO 24612 был подготовлен Техническим комитетом ISO/TC 37, *Терминология и другие языковые и информационные ресурсы,* Подкомитетом SC 4, *Управление языковыми ресурсами*.

Введение

Эффективные процедуры создания, кодирования, обработки языковых ресурсов и управления ими значительно упрощаются при наличии единой высокоуровневой модели данных, которая обеспечивает возможность анализа и проектирования как различных схем аннотирования, так и разнообразных форматов представления аннотаций. Настоящий Международный стандарт предназначен для технической поддержки разработки и использования компьютерных приложений, основой которых служат языковые ресурсы с лингвистическими аннотациями и процедуры обмена такими ресурсами между различными прикладными системами.

Управление языковыми ресурсами. Лингвистическая аннотационная система (LAF)

1 Область применения

Настоящий международный стандарт содержит определение лингвистической аннотационной системы (LAF), которая предназначена для представления лингвистических аннотаций различных языковых данных, таких как текстовые корпуса, речевые сигналы и видеоданные. Эта аннотационная система состоит из абстрактной модели данных и преобразованных в последовательную форму описаний этой модели на языке XML (XML-сериализаций) для представления аннотаций первичных данных. Сериализация служит базовым форматом, позволяющим устанавливать соответствие между аннотациями, представляемыми в разных форматах.

ПРИМЕЧАНИЕ Вопросы стандартизации категорий лингвистических данных, составляющих содержание аннотаций, рассматриваются в ISO 12620 и других аналогичных международных стандартах.

2 Термины и определения

Для целей данного документа используются термины и определения, представленные ниже.

2.1
первичные данные
primary data
языковая информация, представленная в электронной форме

ПРИМЕРЫ Текст, изображение, речевой сигнал.

Примечание к статье: Как правило, обращение к объектам первичных данных осуществляется по адресам их “местоположения” в электронном файле: например, по адресу области памяти, в которой располагаются символы, составляющие предложение или слово, либо по адресу точки, в которой начинается или заканчивается информация об определённом событии (как в случае аннотации речевого сообщения). Более сложные информационные объекты могут представлять собой список или группы последовательно расположенных или разрозненных элементов первичных данных.

2.2
аннотировать, составлять аннотацию
annotate
добавлять лингвистическую информацию к *первичным данным* (2.1)

2.3
аннотация
annotation, noun
лингвистическая информация, добавленная к *первичным данным* (2.1) и не зависящая от формы их представления

2.4
представление
representation
формат, в котором отображается *аннотация* (2.3) , не зависящий от её содержания

ПРИМЕР формат XML, списковый или скобочный формат, текст с разделителями в виде знака табуляции.

2.5
аннотация сегментирования
segmentation annotation
аннотация (2.3), разграничивающая лингвистические элементы, появляющиеся в *первичных данных* (2.1)

Примечание к статье: К числу таких элементов относятся: (1) неразрывные сегменты (появляющиеся в первичных данных совместно); (2) сегменты более высокого или более низкого уровня, являющиеся составными частями более крупного сегмента (например, сегмент из смежных слов, обычно входящий в состав сегмента предложения); (3) дискретные сегменты (для связывания неразрывных сегментов) и (4) реперы (например, отметки времени), обозначающие определённые позиции в первичных данных. В современной практике аннотирования информация сегментирования может присутствовать, а может и не присутствовать в самом документе, содержащем первичные данные.

2.6
лингвистическая аннотация
linguistic annotation
аннотация (2.3), которая предоставляет лингвистическую информацию о сегментах *первичных данных* (2.1)

ПРИМЕР Морфосинтаксическая аннотация, в которой с каждым сегментом данных ассоциируются некоторая часть речи и некоторая лемма.

Примечание к статье: Идентификатор сегмента как слова, предложения, именной группы и т.п. тоже образует лингвистическую аннотацию. В современной практике аннотирования всюду, где это возможно, сегментация часто сочетается с идентификацией лингвистической роли или характеристик сегмента (например, скобочная запись синтаксических свойств или разграничение слов документа с помощью XML-элемента, который определяет сегмент как слово или как предложение).

2.7
автономная аннотация
stand-off annotation
аннотация (2.3), охватывающая различные слои *первичных данных* (2.1) и сериализуемая в документе, отделённом от документа, который содержит первичные данные

Примечание к статье: Автономные аннотации, связываются с конкретными участками первичных данных посредством адресации соответствующих символьных смещений, элементов и т.п. С одни и тем же первичным документом может быть связано множество документированных автономных аннотаций (например, могут существовать аннотации двух разных частей речи, фигурирующих в аннотируемом тексте).

2.8
аннотационный документ, документированная аннотация
annotation document
документ в формате XML, содержащий *аннотации* (2.3)

2.9
якорь, привязка
anchor
жёсткая неизменная позиция в *первичных данных* (2.1), которые необходимо *аннотировать* (2.2)

Примечание к статье: Способ описания якоря определяется конкретной языковой средой. Например, текстовыми якорями могут быть смещения символов, якорями аудиоданных – сдвиги по времени, якорями видеоинформации – временные сдвиги или указатели кадров, а якорями изображений – системы координат.

2.10
местоположение, участок
region
область *первичных данных* (2.1), определяемая непустым упорядоченным списком *якорей* (2.9)

2.11
исходный артефакт
original artefact
искусственный объект или *аннотация* (2.3), используемые для извлечения *первичных данных* (2.1)

2.12
граф
graph
совокупность узлов (вершин) V(G) и связывающих их рёбер E(G)

2.13
узел, вершина
node
vertex
конечная точка в графе G или точка пересечения его рёбер

Примечание к статье: Термины *узел* и *вершина* используются в настоящем документе как синонимы.

2.14
ребро
edge
упорядоченная пара [u,v] узлов, принадлежащих графу, V(G)

Примечание к статье: Порядок следования узлов определяет ориентацию ребра.

3 Спецификация LAF

3.1 Общий обзор

LAF состоит из следующих компонентов:

— информационной модели лингвистических аннотаций и данных, к которым относятся эти аннотации;

— структурной схемы представления языковых данных и их аннотаций;

— сериализованного XML-описания информационной модели, которое характеризует представленную одним или несколькими ориентированными графами ссылочную структуру аннотаций, ассоциируемых с языковыми данными. Узлы графа могут связываться с конкретными участками первичных данных, а в совокупности с рёбрами могут ассоциироваться с соответствующими признаковыми структурами, которые описывают лингвистические свойства участков первичных данных, относящихся к достижимым узлам.

3.2 Модель данных LAF

Модель данных LAF включает в себя следующие блоки:

a) структурное описание информационного носителя, состоящее из *якорей*, указывающих участки первичных данных и их *местоположение*,

b) *графовой структуры*, образованной узлами, рёбрами и ссылками на конкретные участки, и

c) *аннотационной структуры* для представления содержания аннотации с использованием признаковых структур элементов.

Таким образом, информационная модель аннотаций состоит из ориентированного графа, охватывающего *n*-мерные участки первичных данных, и прочих аннотационных представлений, в рамках которых узлы графа ассоциируются с признаковыми структурами, предоставляющими

контент аннотации. Аннотация считается соответствующей LAF, если её схема изоморфна модели данных LAF или может быть преобразована к ней.

ПРИМЕЧАНИЕ В состав лингвистической аннотационной системы не входят спецификации категорий содержания аннотаций (то есть сущностей соответствующих лингвистических явлений).

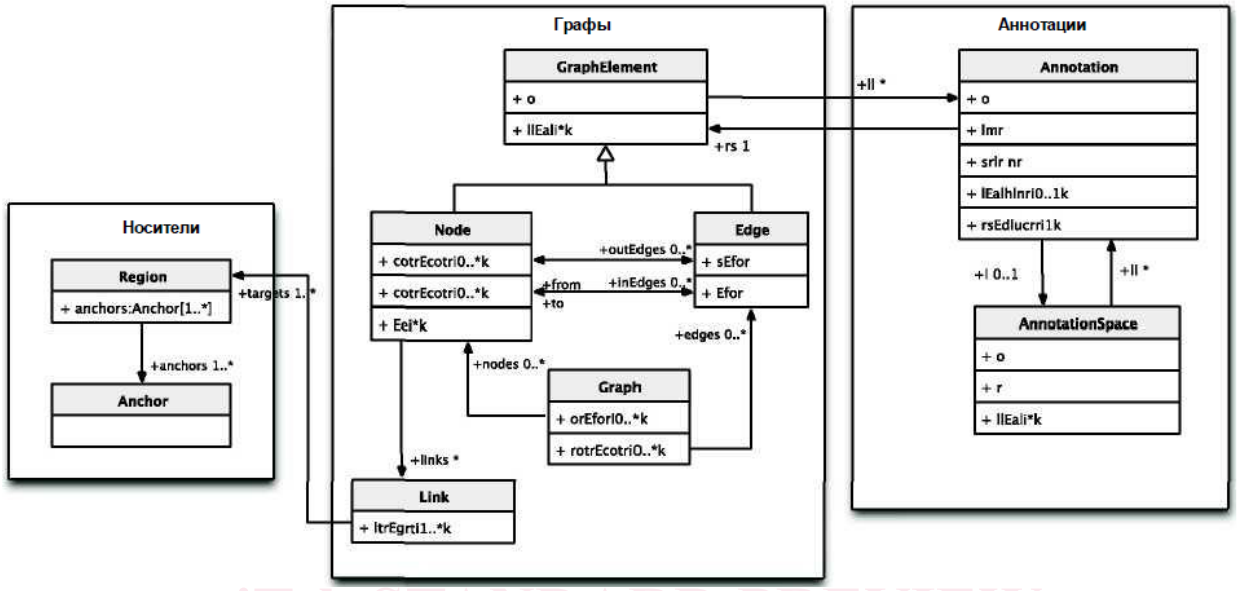

Рисунок 1 — Модель данных LAF

3.3 Архитектура LAF

3.3.1 Общее описание

Языковые ресурсы, соответствующие архитектуре LAF, состоят из перечисленных ниже компонентов, которые более подробно рассматриваются в подразделах 3.3.2 - 3.3.5:

— один или несколько документов, содержащих первичные данные (см. 3.3.2);

— произвольное число документированных аннотаций, охватывающих различные узлы, рёбра графов и ассоциируемые с ними признаковые структуры, все или часть которых могут принадлежать ориентированному графу (орграфу); при этом все узлы снабжаются ссылками либо на базовый документ сегментации (в данном случае узел не имеет исходящих рёбер), либо на другие узлы того же самого или других документов через соответствующие пути в графе (см. 3.3.3);

— один или несколько документов, определяющих области, которые содержат ссылки на каждый документ с первичными данными, служащий основой для сегментации аннотаций (см. 3.3.4.);

— множество заголовочных блоков, включая ресурсный заголовок, описывающий коллекцию документов с первичными данными и аннотациями, равно как и заголовки для каждого первичного документа и каждой аннотации из соответствующей коллекции (см. 3.3.5).

Рекомендуется всегда, когда это возможно, ассоциировать каждый первичный документ с *исходным артефактом*, первичные данные которого извлекаются или адаптируются для аннотации (например, исходный текстовый файл конкретного текстового процессора или программы визуального представления файлов).

3.3.2 Первичные данные

Первичные данные – это сведения, представленные в электронном виде в любом формате, включающие в себя текстовые символы, изображения, аудиоинформацию и видеоданные. Первичные данные в LAF-совместимых ресурсах «замораживаются» как доступные только для чтения (“read-only”) – для обеспечения целостности ссылок на различные участки данных в рамках используемых документов. Внесение корректировок и изменений в первичные данные рассматривается как аннотирование и документируется в отдельной аннотации. Данные текстовых первичных документов имеют кодировку UTF-8 (используемую по умолчанию) или UTF-16.

В общем случае первичные данные не содержат никаких разметочных символов. Если же в первичных данных присутствует разметка (типа тегов HTML или XML), то она воспринимается посредством ссылок на аннотации как часть потока данных; при этом в случае обращения к тем или иным участкам документа не делается никакого различия между символами разметки и символами данных.

3.3.3 Документированные аннотации

Документированная аннотация содержит лингвистическую информацию, предназначенную для описания первичных данных. Аннотации всегда ассоциируются с каким-либо узлом в графе, реализующим прямое обращение к участкам документа, местоположение которых определяется первичными данными непосредственно или по пути, проходящему через достижимые узлы. В последнем случае говорят, что аннотации *расслаиваются* согласно первичным данным. В рамках LAF рекомендуется представлять каждый из лингвистических слоёв, определённых в системе управления языковыми ресурсами, отдельной аннотацией – в целях организации надлежащего информационного обмена.

Степень разбиения аннотации (то есть минимальная единица информации, к которой она применима) зависит от конкретного используемого приложения. Например, единая аннотация некоторого текста может охватывать фонему, слово, предложение, абзац, документ или весь текстовый корпус; а в случае аудиоинформации это могут быть любой временной интервал, включая конкретный «момент времени» (квант времени, временная метка и др.).

3.3.4 Ссылки на первичные данные

Прямое обращение к конкретным участкам первичных данных выполняется с помощью узлов, называемых *якорями*. В большинстве случаев такие узлы располагаются между базовыми единицами представления первичных данных.

Якоря не зависят от характера носителя. Местоположение нужного ресурса может определяться путём задания якорей, ограничивающих участок документа. Участки таких артефактов, как изображение или видеозапись, могут определяться заданием якорей в виде координат местоположения, указателей кадров и т.п. В аудиоданных якоря могут охватывать одну или несколько точек носителя звукозаписи (как, например, “момент” или “временной интервал”). Якоря, осуществляющие такую привязку, представляются комбинациями служебных символов, обозначающих пространственные и временные сдвиги. Например, в английском предложении “My dog has fleas” разбиение может быть произведено так, как показано ниже:

```
                    1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
|M|y| |d|o|g| |h|a|s| |f|l|e|a|s|
```

Здесь со словами связаны следующие якоря:

```
My: начало=0, конец=2
dog: начало=3, конец=6
has: начало=7, конец=10
fleas: начало=11, конец=16
```