МЕЖДУНАРОДНЫЙ СТАНДАРТ

ISO 24611

Первое издание
2012-11-01

Управление языковыми ресурсами. Морфосинтаксическая аннотационная система (MAF)

Language resource management. – Morpho-syntactic annotation framework (MAF)

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

Ссылочный номер
ISO 24611:2012(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 настоящей части стандарта ISO 9735 могут быть объектом патентных прав. ISO не принимает на себя обязательств по определению отдельных или всех таких патентных прав.

ISO 24611 был подготовлен Техническим комитетом ISO/TC 37, *Терминология и другие языковые и информационные ресурсы,* Подкомитетом SC 4, *Управление языковыми ресурсами.*

Введение

Внимание подкомитета SC 4 Технического комитета TC 37 сосредоточено на определении моделей и форм представления аннотированных языковых ресурсов, вследствие чего он распространил стратегию построения моделей, определённую родственным подкомитетом SC 3, на представление терминологических данных [14]; благодаря этому модели лингвистической информации рассматриваются как обобщённые структуры данных (метамодели), которые в дальнейшем детализируются путём отбора соответствующих категорий данных на роль дескрипторов для конкретного уровня аннотирования. Такие модели определяются независимо от каких-либо конкретных форматов и предоставляют в распоряжение специалиста, реализующего определённый проектный замысел, нужный для этого концептуальный инструментарий, который позволяет ему проектировать и сравнивать разные форматы представления по их функциональной эффективности.

Одним из важнейших аспектов представления аннотации любого вида является возможность чёткого и достоверного описания семантики различных используемых дескрипторов – либо в виде формального описания их характеристик и конкретных значений, либо как объектов формализованного представления, например, на языке XML. In order to be shared across various annotation schemas and encoding applications, такие семантические средства должны реализовываться как некий централизованный реестр понятий, к которому пользователь может обращаться как к справочнику категорий данных. Категории данных как таковые должны нести в себе следующие ограничения:

— С технической точки зрения, они должны обеспечивать однозначные стабильные ссылки (реализуемые как постоянные идентификаторы в том смысле, как они определены в ISO 24619), чтобы разработчик конкретной схемы кодирования мог использовать ссылки на стандартизованные категории данных в своём описании. При таком подходе две аннотации будут считаться эквивалентными, когда они определены применительно к одним и тем же категориям (что и характеристика с её значением).

— В дескриптивном плане каждая уникальная семантическая ссылка должна ассоциироваться с подробной документацией, которая содержит в себе полнотекстовый фрагмент описания значения дескриптора с представлением конкретных ограничений, обусловливающих категорию данных.

В последние годы ISO был разработана общая основа для представления и сопровождения такого реестра категорий данных, охватывающего все сферы использования языковых ресурсов. Реализация этой инициативной разработки, описанной в стандарте ISO 12620, привела к созданию оперативно доступной лингвистической среды применительно ко всем категориям данных, которые стандартизуются в рамках многочисленных операций с языковыми ресурсами в связи с внутренней деятельностью ISO, или специально – как часть механизма сопровождения реестра категорий данных. Через этот реестр обеспечивается также доступ к многочисленным категориям данных, которые специалисты по лингвистическим технологиям определяют применительно к конкретным языкам в рамках своей повседневной работы и считают целесообразным довести информацию о них до сведения пользовательского сообщества.

Реестр категорий данных ISO в том виде, как он доступен на сайте ISOCat (www.isocat.org), призван играть роль “не спекулятивной” рыночной площадки семантических объектов, которая накладывает минимум онтологических ограничений. Цель создания подобного реестра заключается в том, чтобы облегчить сопровождение всеобъемлющей дескриптивной среды, в которую легко встраиваются для повторного использования новые категории, без необходимости жёсткой проверки их на соответствие всему реестру в целом. При этом, естественно, частью модели категорий данных являются перечисленные ниже базовые ограничения, как они определены в ISO 12620:

— связи типа “общий - специальный” должны быть простыми, чтобы они могли использоваться для точной идентификации дескрипторов совместимости различных категорий данных. Например, тот факт, что /properNoun/ (имя собственное) является подкатегорией /noun/ (имени существительного), делает возможным сравнение морфосинтаксических аннотаций на основе описаний с разными уровнями детализации;

— описание концептуальных областей должно соответствовать требованиям ISO 11179 для облегчения идентификации возможных значений так называемых сложных категорий данных, когда они применимы или распознаваемы. Например, подобное описание может использоваться для регистрации того факта, что возможные значения концепта /grammaticalGender/ (грамматический род) в малочисленной группе языков [15], могут принадлежать подмножеству {/masculine/, /feminine/ and /neutral/} (мужской, женский и средний);

— ограничения, относящиеся к конкретному языку, должны представляться в форме замечаний по

применению или явно сформулированных ограничений, касающихся концептуальных областей сложных категорий данных. Например, можно в явной форме записать, что концепт /grammaticalGender/ во французском языке может принимать только два значения: {/masculine/ и /feminine/} (мужской и женский).

Настоящий международный стандарт обеспечивает широкую основу для представления аннотаций морфосинтаксических структур (называемых также частями речи). Такая аннотация соответствует первому уровню абстрагирования от лексических значений языковых данных (текстовых или речевых), и в зависимости от языка, в рамках которого осуществляется аннотирование, и от характеристик используемого метода или схемы аннотирования, может в очень широких пределах изменяться по своей структуре и степени сложности.

Для облегчения проработки таких сложных вопросов аннотирования, как обеспечение однозначности и детерминизма, настоящим Международным стандартом определяется метамодель, в которой проводится чёткое различие между двумя уровнями лексических единиц (представляющих сегментацию источника информации на поверхностном уровне) и словоформами (которые идентифицируют лексические абстракции, связанные с группами лексических единиц). Оба этих уровня обладают следующими одинаковыми особенностями: с одной стороны, они могут представляться как простые последовательности и локальные графы множества сегментаций и неоднозначных компоновок, а с другой стороны, все N словоформ могут образовывать комбинации с N лексическими единицами.

В качестве лингвистических сегментов, которые иногда называются в специальной литературе, как, например, в [12], маркерами (‘markables’), *лексические единицы* могут встраиваться в первоисточник информации в виде внутристрочных меток либо могут указывать на него дистанционно посредством так называемых автономных аннотаций.

Словоформы как лингвистические абстракции могут классифицироваться по различным лингвистическим признакам, характеризующим морфосинтаксические свойства, которые приписаны конкретной реализации лексической статьи в рамках аннотируемого текста. Такие свойства могут варьироваться в широком диапазоне – от простого указания на лемму до представленной явным образом ссылки на лексему в словаре. В большинстве существующих приложений морфосинтаксического аннотирования лингвистические характеристики отображаются с помощью так называемых тегов, которые являются кодовым представлением основных признаковых структур (их давние примеры приведены в работе Моначини и Кальзолари [13]). Эти коды могут также нести в себе морфологическую информацию, включая указание части речи (например, существительное, прилагательное или глагол) и такие характеристики, как число, род, лицо, наклонение и глагольное время.

В соответствии с общей стратегией моделирования, принятой Техническим комитетом ISO/TC 37, представленная в настоящем Международном стандарте морфосинтаксическая аннотационная система (MAF) обеспечивает необходимые средства привязки морфосинтаксических тегов, реализуемых признаковыми структурами (согласно ISO 24610), к категориям данных, имеющимся на сайте ISOCat. Нормативное Приложение настоящего Международного стандарта устанавливает множество ключевых категорий данных, которые могут использоваться в режиме ссылок при решении наиболее актуальных текущих задач морфосинтаксического аннотирования в многоязычном контексте. Тем пользователям настоящего Международного стандарта, которые сочтут представленные в нём категории не подходящими им по охвату, сфере применения или семантическим характеристикам, рекомендуется использовать реестр ISOCat для определения собственных категорий данных в соответствии с принципами работы ISO/TC 37.

В соединении с метамоделью MAF обеспечивает также стандартную синтаксическую структуру языка XML, которая может использоваться для сериализации аннотационных моделей, совместимых с MAF. Так как многие существующие лингвистические проекты основываются на рекомендациях Международной организации по кодированию текстовой информации [Text Encoding Initiative (TEI)] (www.tei-c.org), жизненно важных для цифрового представления текстовых первоисточников в компьютеризованном обществе, настоящий Международный стандарт нацелен также на разъяснение способов использования модели MAF в сочетании с TEI-совместимыми методами кодирования. В рамках руководящих указаний TEI уже предложено множество концепций и механизмов для решения широкого круга проблем, связанных с формированием корпусов разговорных языков и их аннотирования [15].

В заключение следует отметить, что данный международный стандарт создаёт концептуальную основу для разработки стандартов серии ISO 24614, касающихся сегментирования текстовой информации, общие принципы и правила которого определены в ISO 24614-1, равно как и для понимания ограничений, излагаемых в дополнительных частях этой серии, которые относятся к конкретным языкам, в соответствии с дихотомией лексема – словоформа.

Управление языковыми ресурсами. Морфосинтаксическая аннотационная система (MAF)

1 Область применения

Настоящий международный стандарт обеспечивает основу для представления аннотаций словоформ в текстах; такие аннотации содержат в себе лексемы, а также их связи с лексическими единицами и морфосинтаксические свойства.

В стандарте описывается метамодель морфосинтаксической аннотации применительно к ссылкам на категории данных, которые содержатся в реестре категорий данных ISOCat (определённом как DCR в ISO 12620). Описывается также сериализация XML-описаний для морфосинтаксических аннотаций в соответствии с рекомендациями TEI (Text Encoding Initiative).

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 24610-1, *Управление языковыми ресурсами. Структуры элементов. Часть 1: Представление структур элементов*

3 Термины и определения

Для целей данного документа используются термины и определения из стандарта ISO 24610-1, а также терминология, приведённая ниже.

3.1
орграф без циклов, ациклический орграф
DAG
directed acyclic graph
граф с ориентированными дугами, не имеющий циклов

Примечание 1 к статье: графы без циклов являются подмножеством *конечных автоматов* (3.4).

3.3
признаковая структура
feature structure
множество спецификаций элементов, используемых в системе морфосинтаксического аннотирования (MAF) для выражения морфосинтаксического содержания

Примечание к статье 1: признаковые структуры описываются как структуры элементов в ISO 24610-1.

3.4
конечные автоматы, КА
FSA
finite state automata
графы переходных состояний, отображающие начальное и конечное состояния и конечное множество переходов автомата из одного состояния в другое

Примечание 1 к статье: см. также *орграф без циклов* (3.1).

3.5
графема
grapheme
минимальная единица письменного языка

ПРИМЕР буква, пиктограмма, идеограмма, число, знак пунктуации.

3.6
изменение формы слова
inflection
модификация или маркировка лексемы, отражающая её морфосинтаксические свойства

3.7
изменённая форма
inflected form
форма, которую слово может принимать в предложении или грамматическом обороте

Примечание 1 к статье: Изменённая форма слова ассоциируется с какой-либо из морфологических характеристик, таких как грамматическое число и падеж.

3.8
лемма
лемматизированная форма
lemma
lemmatised form
общеупотребительная форма представления лексемы

Примечание 1 к статье: В европейских языках лемма обычно представляется в *единственном числе*, если существует множественное; в *мужском* роде, когда существует изменение по *родам*, и в *инфинитиве* глаголов. В некоторых языках определённые имена существительные в форме единственного числа имеют недостаточную парадигму; в таких случаях для представления леммы выбирается множественное число. Для глаголов арабского языка лемма обычно представляется в третьем лице единственного числа совершенного вида.

3.9
лексема
lexeme
морфема, обычно ассоциируемая с множеством словоформ, соответствующих одному общему значению

3.10 лексическая статья
lexical entry
контейнер, обеспечивающий манипулирование множеством словоформ и, возможно, одним или несколькими значениями для описания лексемы

3.11
словарь
lexicon
информационный ресурс, содержащий коллекцию лексических статей некоторого языка

3.12 морфема
morpheme
мельчайшая лингвистическая единица, которая несёт в себе смысл в дискурсе, но не может быть разбита на более мелкие значимые единицы

Примечание 1 к статье: Морфема может быть грамматической (и тогда она называется граммемой) или лексической (т.е. лексемой).

3.13
морфологическая характеристика
морфосинтаксическая характеристика
morphological feature
morpho-syntactic feature
характеристика, выводимая из формы слова

Примечание 1 к статье: Реестр категорий данных ISOCat предоставляет всеобъемлющий список значений для европейских языков.

ПРИМЕР “grammaticalGender” (грамматический род).

3.14
морфология
morphology
описание структуры и способа образования словоформ

ПРИМЕЧАНИЕ 2 Структуры элементов частично упорядочены. Минимальными в этом упорядочении являются пустые структуры элементов.

3.15
морфосинтаксический тег
morpho-syntactic tag
tag
признаковая структура, систематически используемая в качестве спецификатора словоформы

3.16
набор тегов, теговый набор
tagset
исчерпывающее множество тегов, используемых для морфосинтаксического описания языка

Примечание 1 к статье: Для описания набора тегов в качестве ссылочного источника должен использоваться реестр категорий данных ISOCat.

3.17
часть речи
грамматическая категория
part of speech
grammatical category
категория, присваиваемая слову на основе учёта его грамматических и семантических свойств

ПРИМЕР существительное, глагол.

Примечание 1 к статье: Исчерпывающий список значений для частей речи даёт реестр категорий данных ISOCat.

3.18
фонема
phoneme
минимальная единица звуковой системы языка

3.19
типографский шрифт
script
набор графических символов, используемых в одном или нескольких письменных языках

3.20
синтагматическое отношение
syntagmatic relation
отношение, которым связаны лингвистические единицы в дискурсе

3.21
лексический элемент
token
непустая смежная последовательность графем или фонем в документе

Примечание 1 к статье: В некоторых случаях по редакционным причинам в аннотационной схеме понятие лексической единицы может распространяться и на пустую последовательность. См. раздел 6.2, касающийся соединения лексических единиц.

3.22
лексемизация
tokenization
процесс идентификации лексем

3.23
транскрипция
transcription
форма слова, получаемая в результате применения когерентного метода записи звуков речи

3.24
транслитерация
transliteration
форма слова, получаемая в результате преобразования одного текста в другой - обычно путём эквивалентной замены символов

3.25
словоформа
морфосинтаксическая единица
word-form
morpho-syntactic unit
смежная или несмежная лингвистическая единица, определяемая как соответствующая лексическому элементу языка

Примечание 1 к статье: Словоформа может не иметь звуковой или графической реализации либо может соответствовать больше чем одной лексеме.

3.26
словесная решётка
word lattice
совокупность вариантов разбиения сегмента текста или речевого оборота на словоформы

Примечание 1 к статье: Словесная решётка обладает алгебраическими свойствами графа без циклов с начальным и конечным узлами.

Примечание 2 к статье: см. также *орграф без циклов* (3.1) и *конечные автоматы* (3.4).

4 Метамодель MAF

4.1 Общий обзор

Морфосинтаксические аннотации образуют важный уровень лингвистической информации в документе. Настоящий Международный стандарт базируется на метамодели, в которой проводится чёткое разграничение двух уровней: лексических единиц (представляющих поверхностную сегментацию исходного документа) и словоформ (идентифицирующих лексические абстракции, ассоциируемые с группами лексических единиц). У этих двух уровней есть общие особенности: с одной стороны, они могут представляться как простые последовательности и как локальные графы (например, множественными сегментациями и неоднозначными сложными конструкциями), а с другой стороны, все N словоформ могут образовывать комбинации с N лексическими единицами. Настоящий Международный стандарт разграничивает минимальные и максимальные последовательности в документах (текстовые или речевые), могущие идентифицироваться как словоформы, и реализует попытку категоризации лингвистических и распределительных критериев, которые могут использоваться для маркировки этих словоформ в рамках более крупных синтагматических контекстов. Минимальные единицы не могут быть подвергнуты дальнейшему разложению с использованием подобных критериев, но могут быть разбиты на более мелкие единицы с учётом морфологических или фонологических свойств. Словоформы могут агрегироваться в максимальные единицы (такие как сложные или многословные конструкции), которые на других уровнях лингвистического анализа выступают в роли элементарных единиц – в частности, синтаксических. Словоформы могут соответствовать нетерминальному уровню, определённому в ISO 24615.

4.2 Метамодель MAF

Рисунок 1 показывает в упрощённом виде предлагаемую метамодель морфосинтаксических аннотаций, а на Рисунке 2 представлен её более формализованный вид на универсальном языке моделирования UML (Unified Modeling Language).

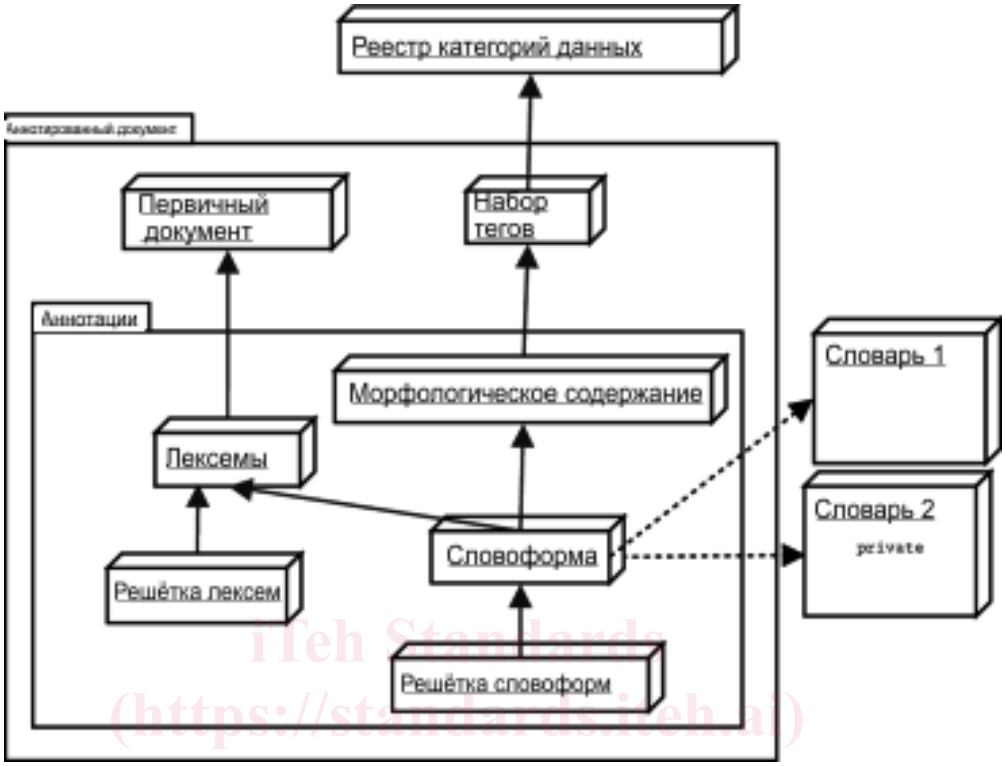

Рисунок 1 — Упрощённый вид метамодели MAF

Аннотированный документ состоит из исходного документа и совокупности аннотаций. Аннотации ассоциируются со словоформами, которые соответствуют нулевому или большему числу лексем в первичном документе. Словоформа может также ассоциироваться с лексической статьёй, которая предоставляет информацию об основополагающей лемме и изменённой форме слова. Морфосинтаксическая аннотация, ассоциируемая со словоформой, представляется тегом, значение которого может быть выражено признаковой структурой. Набор таких тегов, используемый в конкретной схеме аннотирования, называется tagset и в рамках описанных в ISO 24610-2 представлений признаковых структур (FSR) определяется как библиотека структур элементов. Каждая отдельная категория внутри такого набора (tagset) должна допускать описание в терминах каталогизированных категорий данных, представленных в ISO 12620 и реализованных в реестре ISOCat. Поскольку аннотация может составляться применительно как к лексемам, так и к словоформам, возможно появление структурной неоднозначности (омонимии). Поэтому аннотация обычно концептуализируется как одна или несколько ветвей, каждая из которых представляется словесной решёткой или ациклическим орграфом (DAG).

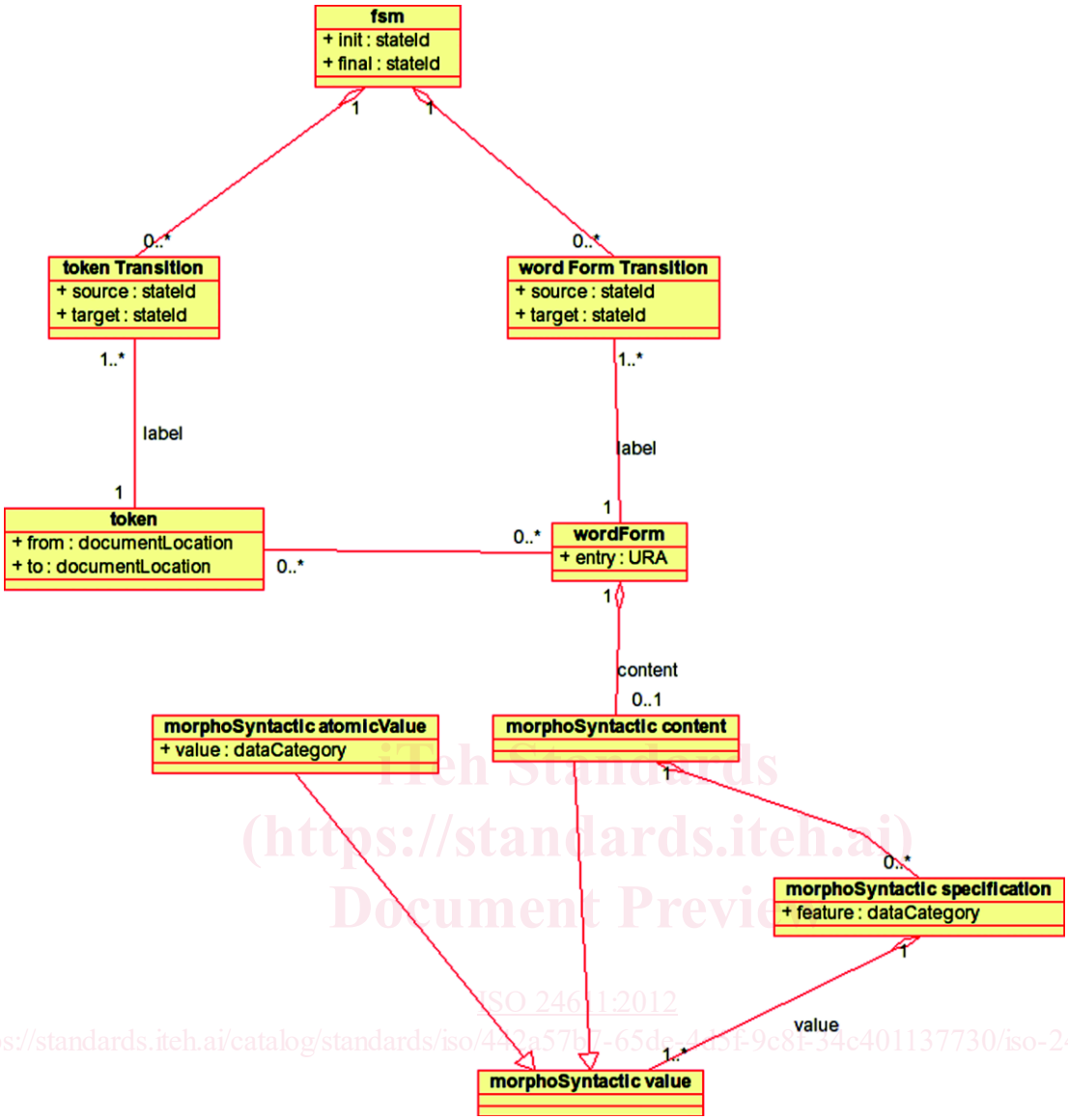

Рисунок 2 — Представление метамодели MAF на языке UML

5 Сегментирование с помощью лексем

5.1 Общие замечания

Морфосинтаксические аннотации строятся на основе сегментов, называемых лексемами, которые присутствуют в потоке документов, но это не означает, что результирующая сегментация будет соответствовать последовательности сопряжённых сегментов, полученных в процессе разбиения первичного документа. Особенно важно проводить различие между словоформами и их конкретными реализациями. Некоторые части документа могут не иметь аннотаций (например, типографские метки, сценические ремарки и элементы разметки), тогда как другие его части могут не иметь точного соответствия сегментированной форме (например, сокращения, скорописные тексты, орфографические ошибки и вариации слов, а также типографские и морфологические контрактуры). Словоформа может не иметь точного соответствия сегменту, который идентифицируется по орфографическим меткам в виде пробелов или дефисов (например, в сложных немецких словах, в транскрипции устной речи и в санскритском письме).

Ниже показаны типичные примеры лексемизированного ввода на двух языках, соответствующие исходному лексическому сегменту, за которым следует представление лексем в виде строки с разделительными символами вертикальной черты: