SLOVENSKI STANDARD
oSIST ISO/DIS 24617-15:2024

01-november-2024

Upravljanje jezikovnih virov - Ogrodje za semantično označevanje (SemAF) - 15. del: Ekstrakcija merljivih kvantitativnih informacij (MQIE)

Language resource management — Semantic annotation framework (SemAF) — Part 15: Measurable quantitative information extraction (MQIE)

Gestion des ressources linguistiques — Cadre d'annotation sémantique (SemAF) — Partie 15: Extraction d'informations quantitatives mesurables (MQIE)

Ta slovenski standard je istoveten z: **ISO/DIS 24617-15**

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
01.140.20	Informacijske vede	Information sciences
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

oSIST ISO/DIS 24617-15:2024 **en,fr**

DRAFT International Standard

ISO/DIS 24617-15

Language resource management — Semantic annotation framework (SemAF) —

Part 15: Measurable quantitative information extraction (MQIE)

ICS: 01.020

ISO/TC **37**/SC **4**

Secretariat: **KATS**

Voting begins on:
2024-08-13

Voting terminates on:
2024-11-05

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 24617-15:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology,* Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24617 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Measurable quantitative information (MQI) describes one of basic properties that is associated with the magnitude aspect of quantity, and is very common in ordinary language. The main characteristics of MQI, as described in ISO 24617-11,[1] is that quantitative information is presented as measures expressed in terms of a pair of a numerically expressed quantity and a unit. Such information is much more abundant in scientific publications or technical reports to the extent that it constitutes an essential part of communicative segments of language in general. The processing of such information is thus required for any successful language resource management.

In such a big data era, demands from industry and academic communities for an accurate extraction of MQI have increased.[2] For example, business investment companies frequently need to identify and aggregate various information covering net sales, gross profit, operating expenses, operating profit, interest expense, net profit before taxes, net income, etc., of the target companies from their annual reports. The fast-growing medical informatics research also needs to process a large amount of medical text to analyze the dose of medicine, the eligibility criteria of clinical trial, the phenotype characters of patients, the lab tests in clinical records, etc.[3,4] All these demands either in industry or in medical research require the effective extraction of MQI for automated identification, aggregation, computation, and analysis [5].

However, in the information retrieval and natural language processing areas, there is no standardized way of extracting measurable quantitative information currently available. Each application system developed in industrial sectors has hitherto used common NLP models or their own models to identify measurable quantitative information from unstructured text. There is no standard extraction procedure for ensuring the quality of the extraction currently. A general, interoperable and standardized measurable quantitative information extraction scheme for IR and NLP tasks to work with many different application systems is called for.

This document, named 'SemAF-MQIE', aims at formulating a general extraction scheme with following the basic requirements of semantic annotation laid down in ISO 24617-11, which facilitates the annotation of MQI in scientific and technical language and to make it interoperable with other semantic annotation schemes such as ISO 24617. The extraction scheme also utilizes various ISO standards on lexical resources and morpho-syntactic annotation frameworks. It aims at being compatible with other existing relevant standards.

NOTE ISO 24617-11 has proposed a standardized schema of annotating measurable quantitative information from unstructured text.

Focusing on measurements in scientifico-technological language, this document is expected to contribute to information retrieval (IR), question answering (QA), text summarization (TS), and other natural language processing (NLP) applications [6-8].

Language resource management — Semantic annotation framework (SemAF) —

Part 15: Measurable quantitative information extraction (MQIE)

1 Scope

This document covers the extractions of measurable or magnitudinal aspect of quantity so that it can focus on the technical or practical use of measurements in IR (information retrieval), QA (question answering), TS (text summarization), and other NLP (natural language processing) applications. It is applicable to the domains of technology that carry more applicational relevance than some theoretical issues found in the ordinary use of language.

NOTE ISO 24617-12 deals with more general and theoretical issues of quantification and quantitative information.

This document also treats temporal durations that are discussed in ISO 24617-1, and spatial measures such as distances that are treated in ISO 24617-7, while making them interoperable with other measure types. It also accommodates the treatment of measures or amounts that are introduced in ISO 24617-6:2016, 8.3.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 24617-11:2021, *Language resource management — Semantic annotation framework (SemAF) — Part 11: Measurable quantitative information (MQI)*

ISO 80000-1:2009, *Quantities and units — Part 1: General*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24617-6:2016, ISO 24617-11:2021 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
information extraction
IE
extracting specific structured information from natural language and/or semi-structured texts and other electronically represented text sources

3.2
measurable quantatitive information extraction
MQIE
extracting measurable quantatitive information from natural language and/or semi-structured texts and other electronically represented text sources

3.3
normalization
process that represents objective information with a formal and/or regular format or converts the information into a consistent value range

Note 1 to entry: The normalization objectives may contain information of entities, measure units and quantities.

4 General framework of MQIE

4.1 Overview

The MQIE usually contains four types of strategies including: 1) manual extraction strategy, 2) semi-automated extraction strategy, 3) automated extraction strategy, and 4) hybrid extraction strategy. The automated extraction strategy usually involves rule-based methods, other machine learning-based methods, and deep learning-based methods. The SemAF-MQIE document thus includes the four types of extraction strategies and specifies a general framework for automated MQIE.

4.2 Primary requirements of MQIE

MQIE shall adhere to the following requirements:

a) Generalable: the extraction process shall be general and adaptive to most kinds of MQI extraction tasks with necessary but slight modification or tuning;

b) Independenable: the extraction process shall not depend on special domains, subjects, and languages;

c) Completable: the extraction process shall be able to identify all elements and links of MQI properly from text by using information extraction related technolgies;

d) Normalizable: International System of Units, decimalism and other metric system shall be adapted to nomalize the extracted quantities, values, and units;

e) Compatible: the extraction result shall be structured, and able to transfor into MQI-compliance formats unambiguously;

f) Assessiable: the extraction result shall be evaluated by widely accepted metrics or criterions;

g) Interpretable: the extraction process and the result shall be explained reasonably upon requests.

4.3 Framework

The overall framework of MQIE is represented by the workflow in Figure 1.