Chapter 10
Migrating from ISO/IEC 9126 to SQUARE:
A Case Study on the Evaluation of Medical Speech Translation Systems

Paula Estrella
Universidad Nacional de Córdoba, Argentina

Nikos Tsourakis
University of Geneva, Switzerland

ABSTRACT

When it comes to the evaluation of natural language systems, it is well acknowledged that there is a lack of common evaluation methodologies, making the fair comparison of such systems a difficult task. Many attempts to standardize this process have used a quality model based on the ISO/IEC 9126 standards. The authors have also used these standards for the definition of a weighted quality model for the evaluation of a medical speech translator, showing the relative importance of the system's features depending on the potential user (patient or doctor, developer). More recently, ISO/IEC 9126 has been replaced by a new series of standards, the 25000 or SQuaRE series, indicating that the model should be migrated to the new series in order to maintain compliance adherence to current standards. This chapter demonstrates how to migrate from ISO/IEC 9126 to ISO 25000 by using the authors' previous work as a use case.

INTRODUCTION

Normalizing evaluations for specific contexts of use have gained increasing importance as software systems become widespread among global organizations and professionals. In particular, Natural Language Processing systems (NLP) are varied in nature and purpose, ranging from low-level applications intended to be used by developers (such as part-of-speech taggers, syntactic parsers or morphological analyzers) to complex systems targeting human end-users (such as voice-commanded booking systems (Jiao et al., 2015) or eye-commanded interfaces (Soltani et al., 2016). During the development lifecycle,

DOI: 10.4018/978-1-5225-1724-5.ch010

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

these systems are periodically evaluated in order to assess the level of improvement achieved, to detect, classify and recover errors or to quantify user acceptability and satisfaction. While, in the last decades numerous authors have provided evaluation results leveraging various computer and human centered metrics, it is well acknowledged that there is a lack of a methodology that would provide a fair comparison framework for different NLP systems.

Convinced that user needs and the specific context of use of NLP systems cannot be omitted in an evaluation, several initiatives emerged which decompose quality into several dimensions. The International Standards for Language Engineering (Calzolari et al., 2002) project was one of these initiatives, which aims at standardizing the evaluation of language engineering systems by relating a customized quality model based on the ISO standards 9126 (ISO/IEC, 2001) to the purpose and context of use of the system based on the ISO standard 14598 (ISO/IEC, 1999). According to these standards, software quality results in general from six categories of quality characteristics (namely functionality, reliability, usability, efficiency, maintainability and portability) that can be particularized to a given software domain and context of use; in that case such a hierarchy is called a quality model and its terminal nodes must be features of the software that can be measured using one or more metrics. Additionally, ISO proposes models to evaluate internal or external quality as well as quality in use.

These standards have been recently replaced by the new 25000 series (ISO, 2014), named SQuaRE, implying that quality models based on the 9126 standard are outdated and should be migrated to the new series. This chapter builds on previous work, where the authors applied the 9126 series to the evaluation of NLP systems in the specific area of medical speech translation systems from the perspective of doctors, patients and developers (Tsourakis & Estrella, 2013). The objective of this chapter is to propose a mapping from a previous weighted quality model to a new weighted quality model based on SQuaRE in order to reuse as much as possible from the previous evaluations, given the complex and laborious work that entails the application of international standards.

BACKGROUND

Quality Models

As quality is hard to assess and assure, several models try to address software quality issues by employing a set of quality attributes, characteristics and metrics. The Factors-Criteria-Metrics (FCM) model was one of the first proposals, extensively used in large projects in the military, space, and public domain (McCall et al., 1977). The model is described in three parts: the first part presents the quality model itself, the second part is devoted to metrics and the third part describes the evaluation process for managers. The FCM model seems more suitable to be applied at later stages of software development cycle, when at least a prototype exists. A similar model was proposed by Boehm et al. (1978), which had a four levels centered on general utility but, despite the terminology, this model is quite similar to the FCM model in structure and characteristics included. Later, the FURPS model, named after the first letters of its top-level quality characteristics, functionality, usability, reliability, performance and supportability, was developed viewing quality as the conformance of a software product to user requirements and to other requirements indirectly obtained, for example from contracts, standards, etc. (Grady, 1992). A more recent proposal took as its starting point the draft version of the ISO/IEC 9126 standard, from which

17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/migrating-from-isoiec-9126-to-square/169549

Related Content

Design Frameworks for Mobile Health Technology: A State-of-the-Art Review of Research From 2015-2021
Ke Zhangand Ayse Begum Aslan (2022). *International Journal of Health Systems and Translational Medicine (pp. 1-13).*
www.irma-international.org/article/design-frameworks-for-mobile-health-technology/302653

Principles of Binocular Stereoscopic Imaging
Geoff Ogram (2018). *Ophthalmology: Breakthroughs in Research and Practice (pp. 78-97).*
www.irma-international.org/chapter/principles-of-binocular-stereoscopic-imaging/195763

Racially Motivated Police Brutality Is a Community Public Health Issue in the United States
Darrell Norman Burrell, Sharon L. Burtonand Grace E. McGrath (2023). *International Journal of Health Systems and Translational Medicine (pp. 1-15).*
www.irma-international.org/article/racially-motivated-police-brutality-is-a-community-public-health-issue-in-the-united-states/315296

Isocenter Verification in Radiotherapy Clinical Practice Using Virtual Simulation: An Image Registration Approach
George K. Matsopoulos, Pantelis A. Asvestas, Vasiliki Markaki, Kalliopi Platoniand Vasilios Kouloulias (2017). *Medical Imaging: Concepts, Methodologies, Tools, and Applications (pp. 1703-1724).*
www.irma-international.org/chapter/isocenter-verification-in-radiotherapy-clinical-practice-using-virtual-simulation/159782

A Survey of Unsupervised Learning in Medical Image Registration
Xin Songand Huan Yang (2022). *International Journal of Health Systems and Translational Medicine (pp. 1-7).*
www.irma-international.org/article/a-survey-of-unsupervised-learning-in-medical-image-registration/282701