Chapter X
Standardization of Terms Applying Finite-State Transducers (FST)

Carmen Galvez
University of Granada, Spain

ABSTRACT

This chapter presents the different standardization methods of terms at the two basic approaches of nonlinguistic and linguistic techniques, and sets out to justify the application of processes based on finite-state transducers (FST). Standardization of terms is the procedure of matching and grouping together variants of the same term that are semantically equivalent. A term variant is a text occurrence that is conceptually related to an original term and can be used to search for information in a text database. The uniterm and multiterm variants can be considered equivalent units for the purposes of automatic indexing. This chapter describes the computational and linguistic base of the finite-state approach, with emphasis on the influence of the formal language theory in the standardization process of uniterms and multiterms. The lemmatization and the use of syntactic pattern-matching, through equivalence relations represented in FSTs, are emerging methods for the standardization of terms.

INTRODUCTION

The purpose of a information retrieval system (IRS) consists of retrieving, from amongst a collection of documents, those that respond to an informational need, and to reorganize these documents according to a factor of relevance. This process normally involves *statistical methods* in charge of selecting the most appropriate terms for representing documental contents, and an *inverse index file* that accesses the documents containing these terms (Salton & McGill, 1983). The relationship of pertinence between queries and documents is established by the number of

terms they have in common. For this reason the queries and documents are represented as sets of characteristics or indexing terms, which can be derived directly or indirectly from the text using either a thesaurus or a manual or automatic indexing procedure. In many IRS, the documents are indexed by uniterms. However, these may result ambiguous, and therefore unable to discriminate only the pertinent information. One solution to this problem is to work with multiword terms (or *phrases*) often obtained through statistical methods. The traditional IRS approach is based on this type of automatic indexing technique for representing documentary contents (Croft, Turtle, & Lewis, 1991; Frakes, 1992; Salton, 1989).

Matching query terms to documents involves a number of advanced retrieval techniques, and one problem that has not yet been solved is the inadequate representation of the two (Strzalkowski, Lin, Wang, & Pérez-Carballo, 1999). At the root of this problem is the great variability of the lexical, syntactic, and morphological features of a term, variants that cannot be recognized by simple *string-matching algorithms* without some sort of *natural language processing* (NLP) (Hull, 1996). It is generally agreed that NLP techniques could improve IRS yields; yet it is still not clear exactly how we might incorporate the advancements of computational linguistics into retrieval systems. The grouping of morphological variants would increase the average recall, while the identification and grouping of syntactic variants is determinant in increasing the accuracy of retrieval. One study about the problems involved in using linguistic variants in IRS is detailed by Sparck Jones and Tait (1984).

The term standardization is the process of matching and grouping together variants of the same term that are semantically equivalent. A variant is defined as a text occurrence that is conceptually related to an original term and can be used to search for information in text databases (Jacquemin & Tzoukermann, 1999; Sparck Jones & Tait, 1984; Tzoukermann, Klavans, & Jacquemin, 1997). This is done by means of computational procedures known as *standardization or conflation algorithms*, whose primary goal is the normalization of uniterms and multiterms (Galvez, Moya-Anegón, & Solana, 2005). In order to avoid the loss of relevant documents, an IRS recognizes and groups variants by means of so-called conflation algorithms. The process of standardization may involve linguistic techniques such as the segmentation of words and the elimination of affixes, or lexical searches through thesauri. The latter is concerned with the recognition of semantic variants, and remains beyond the scope of the present study.

This chapter focuses on the initial stage of automatic indexing in natural language, that is, on the process of algorithmically examining the indexing terms to generate and control the units that will then be incorporated as potential entries to the search file. The recognition and grouping of lexical and syntactic variants can thus be considered a process of normalization; when a term does not appear in a normalized form, it is replaced with the canonical form. Along these lines, we will review the most relevant techniques for grouping variants, departing from the premise that conflation techniques featuring linguistic devices can be considered normalization techniques, their function being to regulate linguistic variants.

THE PROBLEM OF TERM VARIANTS

During the first stage of automatic indexing in natural language we encounter a tremendous number of variants gathered up by the indexing terms. The variants are considered semantically similar units that can be treated as equivalents in IRS. To arrive at these equivalencies, standardization methods of variants are used, grouping the terms that refer to equivalent concepts. The variants can be used to extract information in the textual databases (Jacquemin & Tzoukermann,