Linguistic Knowledge Discovery from Character Encoding Properties

Dafydd Gibbon[1], Baden Hughes[2], and Thorsten Trippel[1]

[1] Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld, Postfach 100 131, D-33501 Bielefeld, Germany
[2] Department of Computer Science and Software Engineering
University of Melbourne, Parkville 3010, Australia

Abstract. While much focus on linguistic knowledge is at the level of data analysis, an inherent, but little explored domain rich in linguistic information are the character sets and encodings with which linguistic data is represented. Despite the increasing prevalence of Unicode as a character encoding standard, there remains a large volume of linguistic data which is encoded using legacy character encodings and fonts. For the most part, this data is not systematised at the character level, either internally or with reference to an external standard. In order to unify the inherently linguistic properties of both best-practice and legacy character encodings, we require an analytical and representational approach independent of a font or character encoding itself. We reduce a character to a series of linguistic feature vectors, (such as phonetic, phonemic and orthographic characteristics), complemented by the inheritance a range of properties from Unicode (such as inherent directionality, combining behaviour etc). This resulting analysis model is applicable both to Unicode and non-Unicode character encodings, and hence provides a leverage point for the discovery of linguistically- grounded character information independent of a specific font. Having establised an analytical and representational mechanism, we can proceed to classify characters using an similarity matrix, and discover linguistic knowledge as to their properties. At a basic level, individual characters and sequences of characters can be classified as similar or different from a number of perspectives including their relative proximity in binary space, their linguistic and rendering behaviours in relation to a given context, their linguistic similarity, their descriptive similarity, provenance throughout a family of related fonts etc. Naturally, these metrics can be displayed in a number of ways for interpretation. In order to manage these classifications, we induce an ontological approach - considering character encodings used within a single font as a type of namespace, and relating many different encodings to a single set of common units. The analytical and representational model presented here allows us to conduct a range of data mining operations over linguistic data regardless of its character encoding expression. Furthermore, linguistic properties are able to be used as query terms. We offer this technique as a contribution to enabling linguistic knowledge discovery in a new context.

Keywords

character encoding, font ontologies, semantic decomposition, linguistic knowledge