

ABSTRACT

Embodiments of the present application provide a cross-language data enhancement-based word segmentation method and apparatus. The technical solution provided in the embodiments of the present application is: acquiring high-resource language data and processing same to obtain word segment materials, acquiring low-resource language data to obtain candidate word segments, and screening the candidate word segments according to the word segment materials obtained from the high-resource language data to select candidate word segments having a high degree of matching with the word segment materials as word segment materials of the low-resource language data, training a word segmentation model on the basis of the low-resource word segment materials to automatically output candidate word segmentation results for the low-resource language data on the basis of the model, and selecting a word segmentation result according to the degree of matching between the candidate word segmentation results and the word segment materials of the high-resource language data. The use of high-resource language materials to automatically augment and validate model training data in low-resource languages solves the problem of imbalance between data resources and annotation resources in different languages, thereby providing an easier and more efficient solution for an iterative word segmentation model.

FIG. 1