Abbreviation Recognition with MaxEnt Model

Chunyu Kit, Xiaoyue Liu, and Jonathan J. Webster

Department of Chinese, Translation and Linguistics
City University of Hong Kong, 83 Tat Chee Ave., Kowloon, Hong Kong
{ctckit,xyliu0,ctjjw}@cityu.edu.hk

Abstract. Abbreviated words carry critical information in the literature of many special domains. This paper reports our research in recognizing dotted abbreviations with MaxEnt model. The key points in our work include: (1) allowing the model to optimize with as many features as possible to capture the text characteristics of context words, and (2) utilizing simple lexical information such as sentence-initial words and candidate word length for performance enhancement. Experimental results show that this approach achieves impressive performance on the WSJ corpus.

1 Introduction

The literature in many special domains, e.g., biomedical, has been growing rapidly in recent years with a large number of abbreviations carrying critical information, e.g., proper names and terminology. There is an increasing interest in practical techniques for identifying abbreviations from plain texts.

There are several typical forms of abbreviation, including acronyms, blending, and dotted strings. Previous research [2, 7] illustrated significant success in identifying and pairing short form terms, referred to as abbreviations, most of which are acronyms, and their original long forms, referred to as definitions, e.g., `<HIV, Human Immunodeficiency Virus>`. This paper is intended to report our recent work to apply the Maximum Entropy (MaxEnt) model to identify abbreviations in another form, i.e., dotted strings, e.g., “abbr.” for “abbreviation”, “Jan.” for “January”. Popular abbreviations of this kind such as “Mr.”, “Dr.”, “Prof.”, “Corp.” and “Ltd.” are available from an ordinary dictionary. There is no point to invent any complicated techniques for recognizing them. The availability of such a sample set, however, gives us great convenience to evaluate the performance of a learning model on recognizing abbreviations with a particular surface form. The significance of this approach lies in the plausibility that similar methodology can be applied to abbreviations with some other common surface characteristics, e.g., in parentheses.

Aiming at this objective, we intend to allow the MaxEnt model to optimize with as many features as possible to capture the text form characteristics of context words and with some special features to utilize simple lexical information such as candidate word length and sentence-initial words that can be derived from the training data straightforwardly. Section 2 below presents feature selection for MaxEnt model training, and Sect. 3 the experiments for evaluation. Section 4 concludes the paper in terms of experimental results.