Periods, Capitalized Words, etc.

Andrei Mikheev*
University of Edinburgh

In this article we present an approach for tackling three important aspects of text normalization: sentence boundary disambiguation, disambiguation of capitalized words in positions where capitalization is expected, and identification of abbreviations. As opposed to the two dominant techniques of computing statistics or writing specialized grammars, our document-centered approach works by considering suggestive local contexts and repetitions of individual words within a document. This approach proved to be robust to domain shifts and new lexica and produced performance on the level with the highest reported results. When incorporated into a part-of-speech tagger, it helped reduce the error rate significantly on capitalized words and sentence boundaries. We also investigated the portability to other languages and obtained encouraging results.

1. Introduction

Disambiguation of sentence boundaries and normalization of capitalized words, as well as identification of abbreviations, however small in comparison to other tasks of text processing, are of primary importance in the developing of practical text-processing applications. These tasks are usually performed before actual "intelligent" text processing starts, and errors made at this stage are very likely to cause more errors at later stages and are therefore very dangerous.

Disambiguation of capitalized words in mixed-case texts has received little attention in the natural language processing and information retrieval communities, but in fact it plays an important role in many tasks. In mixed-case texts capitalized words usually denote proper names (names of organizations, locations, people, artifacts, etc.), but there are special positions in the text where capitalization is expected. Such mandatory positions include the first word in a sentence, words in titles with all significant words capitalized or table entries, a capitalized word after a colon or open quote, and the first word in a list entry, among others. Capitalized words in these and some other positions present a case of ambiguity: they can stand for proper names, as in *White later said* . . . , or they can be just capitalized common words, as in *White elephants are* . . . . The **disambiguation of capitalized words** in ambiguous positions leads to the **identification of proper names** (or their derivatives), and in this article we will use these two terms and the term **case normalization** interchangeably.

Church (1995, p. 294) studied, among other simple text normalization techniques, the effect of case normalization for different words and showed that "sometimes case variants refer to the same thing (*hurricane* and *Hurricane*), sometimes they refer to different things (*continental* and *Continental*) and sometimes they don't refer to much of anything (e.g., *anytime* and *Anytime*)." Obviously these differences arise because some capitalized words stand for proper names (such as *Continental*, the name of an airline) and some do not.

* Institute for Communicating and Collaborative Systems, Division of Informatics, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: mikheev@cogsci.ed.ac.uk