LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Mining Wikipedia as a Parallel and Comparable Corpus

JESÚS TOMÁS
JORDI BATALLER
AND FRANCISCO CASACUBERTA
Instituto Tecnológico de Informática

JAIME LLORET
Universidad Politécnica de Valencia

ABSTRACT

It is difficult to find parallel text corpora but for a few languages or for specific domains. Recently, collaborative edited multilingual projects, like Wikipedia, are becoming widespread. This paper studies the feasibility of using Wikipedia to obtain parallel corpora. Two types of articles can be used: parallel and comparable. We explore a distinct approach for each type of article. Parallel articles are identified by using – language independent – Web Mining techniques. For comparable articles, we use a classifier. It is based on a log-lineal combination of feature-functions, tuned with minimum error criterion. In order to validate our approaches, we report two experiments. First, a restricted domain corpus (pharmacology) in two great diffusion languages (English and Spanish) is obtained. In the second, the corpus is generated for a minority language (Catalan) and Spanish.

1. INTRODUCTION

Inductive methods are currently being considered with increasing interest in multilingual computational linguists. They have an important limitation: a parallel corpus is needed to train their models. For instance, a parallel corpus is required in tasks such as machine translation (Brown et al. 1993; Och & Ney 2000), cross-lingual information retrieval (Chen and Nie 2000) or lexical acquisition (Brown et al. 1991). There are few parallel corpora, for a few languages, and often in restricted domains. Currently, we can find large parallel