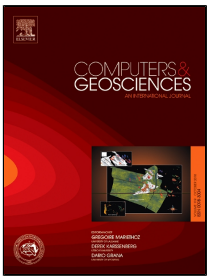
Accepted Manuscript



DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain

Qinjun Qiu, Zhong Xie, Liang Wu, Wenjia Li

PII:	S0098-3004(18)30085-2
DOI:	10.1016/j.cageo.2018.08.006
Reference:	CAGEO 4169
To appear in:	*Computers and Geosciences*
Received Date:	27 January 2018
Accepted Date:	31 August 2018

Please cite this article as: Qinjun Qiu, Zhong Xie, Liang Wu, Wenjia Li, DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain, *Computers and Geosciences* (2018), doi: 10.1016/j.cageo.2018.08.006

DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain

Qinjun Qiu[1, 2], Zhong Xie[1,2], Liang Wu[1,2, *], Wenjia Li[1,2]
qiuqinjun@cug.edu.cn; xiezhong@cug.edu.cn; cugqqj@163.com; liwenjia@cug.edu.cn
[1] Department of Information Engineering, China University of Geosciences, Wuhan 430074, China
[2] National Engineering Research Center of Geographic Information System, Wuhan 430074, China

Abstract: Larger numbers of geoscience reports create challenges and opportunities for data analysis and knowledge discovery. Segmenting texts into semantically and syntactically meaningful words is known as the Chinese word segmentation (CWS) problem because there is no space between words in the Chinese language. CWS is a crucial first step toward natural language processing (NLP). Although the available generic segmenters can process geoscience reports, their performance degrades dramatically without sufficient domain knowledge. Hence, developing effective segmenters remains a challenge and requires more work.
This inspired us to build a segmenter for the geoscience subject domain. By integrating the unigram language model and deep learning, we propose a weakly supervised model: DGeoSegmenter. DGeoSegmenter is trained with words and corresponding frequencies. We built DGeoSegmenter using the bi-directional long short-term memory (Bi-LSTM) model, which randomly extracts words and combines them into sentences. Our evaluation results using geoscience reports and benchmark datasets demonstrate the effectiveness of our method, DGeoSegmenter can segment both geoscience terms and general terms. Since manually labeled datasets and hand-crafted rules are not necessary for this proposed algorithm, it can easily be applied to various domains including information retrieval and text mining.

Keywords: Chinese word segmentation Geoscience reports Unigram language model Natural language processing

1 Introduction

The explosive growth of geological reports has caused their accumulation during geological survey procedures. The reports include various geological topics, such as rocks, minerals, and hydrology. In addition, large amounts of unstructured data are difficult to manage and store via virtual applications. For unstructured geological data, they contain more abundant information and have more potential value than structured data (Wu et al., 2017). In recent years, considerable research on mathematical geoscience efforts has been devoted to discovering new knowledge about georeferenced

* Corresponding author: Liang Wu. Conceived and designed the experiments: Qinjun Qiu, Liang Wu, and Zhong Xie; Performed the experiments: Qinjun Qiu, Liang Wu, and Zhong Xie; Analyzed the data: Qinjun Qiu, Liang Wu, and Zhong Xie; Wrote the paper: Qinjun Qiu, Liang Wu, Zhong Xie and Wenjia Li.