Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 48–57

Procedia
Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

Resolving Entity Morphs based on Character-Word Embedding

Ying Sha[1,2], Zhenhui Shi[1,2], Rui Li[1,2], Qi Liang[1,2], and Bin Wang[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences (CAS), Beijing, China
{shaying,shizhenhui,lirui,liangqi,wangbin}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences.

Abstract

Morph is a special type of fake alternative names. Internet users use morphs to achieve certain goals such as expressing special sentiment or avoiding censorship. For example, Chinese internet users often replace “马景涛”(Ma Jingtao) with “咆哮教主”(Roar Bishop)[1]. “咆哮教主”(Roar Bishop) is a morph and “马景涛”(Ma Jingtao) is the target entity of “咆哮教主”(Roar Bishop). This paper focuses on morph resolution: given a morph, figure out the entity that it really refers to. After analyse the common characteristic of morphs and target entities from cross-source corpora, we exploit temporal and semantic constraints to collect target candidates. We propose a framework based on character-word embeddings and radical-character-word embeddings to rank target candidates. Our method does not need any human-annotated data. Experimental results demonstrate our approaches outperforms the state-of-the-art method. The results also show that the performance is better when morphs share any character with target entities.

Keywords: morph; morph resolution; social network; word embedding; character-word embeddings

1 Introduction

Many social media users create morphs [1] , the fake alternative names, to entertain readers or avoid internet censorship. Morphs are widely used in Chinese social media. For example, there was a piece of Chinese Sina Weibo tweet:”Little Brother Ma (小马哥) is not a small fresh meat now. He can also be behind closed doors in the island. But he should adapt to his roles when he goes out”. Here a morph ” 小马哥” (Littler Brother Ma) was created to refer to ” 马英九” (Ma Ying-jeou)[2].

Apparently, the successful resolution of morphs is very important for understanding social media, and it can potentially benefit various online applications, such as information extraction, search engines, automatic question-answering, and recommendation systems.

[1] Ma Jingtao is an actor who usually use exaggerated roaring to express emotions.
[2] Ma Ying-jeou is a former Taiwan leader

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.106

This paper mainly focuses on morph resolution: given a morph, figure out the entity that it really refers to. The morph resolution is very difficult because of the following challenges:

1) The distributions of co-occurrence of morphs and their target entities are quite different in different social media;
2) Most morphs were created according to the semantic links between morphs and their target entities based on historical and cultural narratives [1];
3) Tweets from Twitter or Chinese Sina Weibo are short text and noisy. Therefore it is not easy to extract enough evidences and contexts for morph resolution.

Huang et al. [1] did a pioneering study on morph resolution, but their method used a large amount of human-annotated data and their approach was context-independent. Zhang et al. [2] summarized 548 randomly selected morphs into 8 categories. Zhang et al. [3] proposed the first end-to-end context-aware entity morph decoding system. But they did not take consider of the semantic links of characters of morphs and target entities.

In this paper, we propose a framework based on character-word embeddings and radical-character-word embeddings to explore the semantic links between morphs and target entities. First, we analyse the common characteristics shared among morphs and target entities from temporal distribution, topic similarity, and cross-sources datasets. Then based on these common characteristics, we collect target candidates. Next, we get the semantic meanings of morphs and target candidates based on character-word embeddings and radical-character-word. Finally, we rank target candidates based on similarity measurement of semantic meanings of morphs and target candidates. Using this method, we take consider of both the external semantic meanings (word) and internal semantic meanings (radical and character) of morphs and target candidates.

Our approach does not require any manually constructed morph-target entities pairs for training and outperforms the state-of-the-art method [3]. The results also show that the performance is better when morphs share any character with target entities (The accuracy can reach 87%).

Our paper offers the following contributions:

1) In order to narrow down the scope of target candidates and not leave out real target entities, we set the threshold of temporal slot to 4 days.
2) In order to quickly and efficiently resolve the real target entities, we collect target candidates based on heterogeneous information from multiple sources.
3) we propose a framework based on character-word embeddings and radical-character-word embeddings to address the issue of morph resolution using both external semantic links and internal semantic links.

2 Related Work

Morph resolution is closely related to alias detection [4]. Other similar research include mining name translation pairs from comparable corpora [5] and link prediction [6]. Most of the work focused on unstructured and structured data with clean and rich relations. It has been demonstrated these techniques did not perform well when directly applying on morph resolution.

To our knowledge, Huang et al. [1] proposed the first work on morph resolution, their results have served as a benchmark for this problem. Zhang et al. [2] summarized 548 randomly selected morphs into 8 categories: (1) Phoentic Substitution; (2) Spelling Decomposition; (3) Nickname Generation; (4) Translation and Transliteration; (5) Semantic Interpretation; (6) Historical