Journal of King Saud University – Computer and Information Sciences (2014) 26, 379–387

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Journal of King Saud University – Computer and Information Sciences

Transliteration normalization for Information Extraction and Machine Translation

Yuval Marton, Imed Zitouni [*]

Microsoft, Bellevue, WA, United States

Available online 12 October 2014

KEYWORDS

Arabic;
Named Entity Recognition;
Transliteration;
Name normalization;
Information Extraction;
Machine Translation

Abstract Foreign name transliterations typically include multiple spelling variants. These variants cause data sparseness and inconsistency problems, increase the Out-of-Vocabulary (OOV) rate, and present challenges for Machine Translation, Information Extraction and other natural language processing (NLP) tasks. This work aims to identify and cluster name spelling variants using a Statistical Machine Translation method: word alignment. The variants are identified by being aligned to the same "pivot" name in another language (the source-language in Machine Translation settings). Based on word-to-word translation and transliteration probabilities, as well as the string edit distance metric, names with similar spellings in the target language are clustered and then normalized to a canonical form. With this approach, tens of thousands of high-precision name transliteration spelling variants are extracted from sentence-aligned bilingual corpora in Arabic and English (in both languages). When these normalized name spelling variants are applied to Information Extraction tasks, improvements over strong baseline systems are observed. When applied to Machine Translation tasks, a large improvement potential is shown.

1. Introduction

Foreign names typically have multiple spelling variants after translation or transliteration (where translation aims to preserve meaning, while transliteration aims to preserve sound, given differences in the languages' sounds and writing systems). These spelling variants present challenges for many natural language processing (NLP) tasks, as they increase both the vocabulary size and Out-of-Vocabulary (OOV) rate,[1] exacerbate the data sparseness problem, and may introduce inconsistencies (in spelling or in reference as multiple entities). When different spelling variants are generated for the same name in one document, it reduces the named entity resolution scores and the readability of Machine Translation output. This paper addresses this problem by replacing each spelling variant with a corresponding canonical form. Such text normalization could potentially benefit many NLP tasks, including information retrieval, Information Extraction, question answering, speech recognition, and Machine Translation.

[*] Corresponding author.

[1] OOV rate: how often the model processes an input term that it has not been trained on. Typically, models perform poorly on OOV terms, whether they be Machine Translation, parsing, Mention Detection or other NLP models.

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.011
1319-1578 © 2014 Production and hosting by Elsevier B.V. on behalf of King Saud University.

Name spelling variants have been studied mostly in Information Retrieval (IR) research, especially in query expansion and cross-lingual IR. Bhagat and Hovy (2007) proposed two approaches for (primarily English) spelling variant generation, based on letters-to-phonemes mapping and the SoundEx algorithm (Knuth, 1973). Raghavan and Allan (2005) proposed several techniques to group names in Automatic Speech Recognition (ASR) output and evaluated their effectiveness in spoken document retrieval (SDR). Both approaches use a named entity extraction system to automatically identify names. For multi-lingual name spelling variants, Linden (2006) proposed using a general edit distance metric with a weighted FST to find technical term translations (which were referred to as "cross-lingual spelling variants"). These variants are typically translated words with similar stems in another language. Toivonen and colleagues (2005) proposed a two-step fuzzy translation technique to solve similar problems. Al-Onaizan and Knight (2002), Huang et al. (2003), and Ji and Grishman (2007) investigated the general name entity translation problem, especially within the context of Machine Translation.

All of these approaches rely on name taggers and other classifiers to directly identify the variants. This work, however, aims to identify name spelling variants using *crosslingual* information, with application to Arabic and English. Instead of using a named entity tagger to directly identify names and their spelling variants, we link spelling variants with a name in another language via a method that is widely used in Statistical Machine Translation: word alignment. From sentence-aligned bilingual corpora, we collect word co-occurrence statistics and calculate word translation probabilities (including transliterated words).[2] For each source-side word, we group its target-side aligned counterparts into clusters according to target-side string edit distances. Then, we calculate the transliteration cost between the source word and each target-side cluster (see Section 3). Word pairs with small transliteration costs are considered name variants. We then normalize all names in each cluster to the most frequent form.

Note that spelling variation does not necessarily stem from transliteration or translation, e.g.,

- Cindy and Cyndi
- Kacey and KC (read-aloud initials)
- Cl8n and Clayton (informal communication writing style)
- Dialectal differences (e.g., الجزيرة vs. الڨزيرة)

However, these other cases most likely should not be clustered and normalized (except, perhaps, the informal writing style), as they are likely to refer to different people/entities. These cases are outside the scope of this work.

We applied our approach to extract name transliteration spelling variants from bilingual Arabic–English corpora. We obtained tens of thousands of high-precision name translation pairs. We further applied these spelling variants to Machine Translation (MT) and Information Extraction (IE) tasks, and observed a statistically significant improvement over a strong baseline on the IE task, and a close to "oracle" improvement on a small test set on the MT task.

After an Arabic-focused survey of related work (Section 2), we describe our model setting in both Information Retrieval and Statistical Machine Translation (Section 3). We then detail our past and new experiments (Section 4). We follow up with an analysis of the results (Section 4) and conclude with possible future work (Section 5).

2. Related work

In addition to the work we mentioned earlier, there has been much related work in both IE and MT. We focus here on Arabic (or Arabic and English) related work.

The idea of using cross-language propagation to boost performance has been applied by several researchers. For example, Tackstromand et al. (2012) show how the use of cross-lingual word clusters for the transfer of linguistic structure improves system performance. Other research studies (such as Goldsmith, 2001; McCallum and Nigram, 1998; Yarowsky, 1995) report the use of cross-language propagation to boost the performance of different systems, namely, morphological segmentation, text categorization and word segmentation, respectively. These approaches are based on monolingual data. Rogati et al. (2003) use a Statistical Machine Translation (SMT) system to build an Arabic stemmer. The obtained stemmer has a performance of 87.5%. Ide et al. (2002) use the aligned versions of George Orwell's "Nineteen Eighty-Four" in seven languages to determine sense distinctions that can be used in the Word Sense Disambiguation (WSD) task. They report that the automatically obtained tags are at least as reliable as the tags created by human annotators. Zitouni et al. (2005) attempt to enhance a Mention Detection model of a foreign language by using an English Mention Detection system. They used an SMT system to (i) translate the text into English, (ii) run the English model on the translated text, and (iii) propagate the outcome to the original text. Das and Petrov (2011) try a similar approach but apply it to POS taggers. Both approaches require an SMT system.

The detection (or generation) of named entity variants has also been explored and evaluated in SMT, often as a subset of a paraphrase generation task. In this case, variants (paraphrases) are used to augment translation tables that are missing the variants, unlike our work, which uses them for the normalization of existing terms. Hereafter, we call the term to be paraphrased the *anchor*.

Callison-Burch et al. (2006) proposed a general paraphrasing method by "pivoting" through additional languages in SMT tables and back to the original language. The method is as follows: for each anchor, find its translation(s) in the table and "pivot" through each translation term back to the original (the anchor's) language, i.e., translate back. The back-translations are often good paraphrases and potentially good name variants. Our work uses similar pivoting but then further clusters terms by edit distance and transliteration cost. Interestingly, Callison-Burch et al. (2006) excluded named entities from their experiments, presumably due to noisier results in this particular subset problem. Callison-Burch (2008) improved this method with syntactic constraints. Many publications used or extended the pivoting method, some of which we list below. While the

[2] Throughout this article, we sometimes use the term 'translation' loosely, encompassing both translation and transliteration, as there is no explicit representational difference between the two in Statistical Machine Translation phrase tables.