Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 111–118

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Detecting and Integrating Multiword Expression into English-Arabic Statistical Machine Translation

Sara Ebrahim[1], Doaa Hegazy[1], Mostafa Gadal-Haqq M. Mostafa[2], Samhaa R. El-Beltagy[3]

Abstract

In this paper we introduce a new method for detecting a type of English Multiword Expressions (MWEs), which is phrasal verbs, into an English-Arabic phrase-based statistical machine translation (PBSMT) system. The detection starts with parsing the English side of the parallel corpus, detecting various linguistic patterns for phrasal verbs and finally integrate them into the En-Ar PBSMT system.

In addition, the paper explores the effect of cliticizing specific words in English that have no Arabic equivalent. The results, which reported with the BLEU scores, showed that some patterns achieved significant improvements compared to other patterns and still the baseline achieves the highest score.

This paper shows that, by detecting more linguistic patterns and integrating them into En-Ar SMT system, translation quality could be improved with other integration methods. Yet, the results show which path is worth to follow and clarifies the perspective that linguistic features are not handled properly in the statistically learned models.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Statistical Machine Translation (SMT), Phrase-based SMT, Arabic Natural Language Processing, Multiword Expressions

1. Introduction

Modeling Multiword expressions (MWEs) is a key problem for the development of Natural Language Processing (NLP) fields. Nevertheless, integrating them in large-scale Statistical Machine Translation (SMT) systems has proved to be a promising area, despite being hard to model, especially in distant language pair such as English-Arabic. In the scope of SMT, MWEs are sentences that have a meaning different from the literal meaning for each word in the sentence. More scientifically, MWEs are defined in [14] as ”idiosyncratic concepts that cross word boundaries

E-mail addresses: sara.elkafrawy@gmail.com (Sara Ebrahim)., doaa.hegazy@cis.asu.edu.eg (Doaa Hegazy)., mgmostafa@cis.asu.edu.eg (Mostafa Gadal-Haqq M. Mostafa)., samhaa@computer.org (Samhaa R. El-Beltagy).

[1] Scientific Computing Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
[2] Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
[3] Center of Informatics Science, Nile University, Giza, Egypt

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.099

(spaces)." . It is sometimes an issue for junior human translators to transfer the correct meaning of an MWE from the source language into the target one. Moreover, the amount of research concerning integrating MWEs into MT systems is not significantly enough compared to detecting MWEs.

This paper proposes a new method for detecting and integrating English phrasal verbs as an MWE into English-Arabic phrase-based Statistical Machine Translation System (PBSMT). The method is described in more details in the fourth section. In the second section we list how scientists classified the MWEs types. Although MWE integration into SMT systems was rarely presented, we will list the related work for integrating different kinds of MWEs into En-Ar SMT systems in section three. In section five and six, we evaluate the experiments with BLEU score and discuss the results with a linguistic perspective in order to extend other useful pattern combinations. Finally, we give our conclusion in the last section of this paper.

2. Multiword Expressions Types

MWEs have many types. Gheneim et. al. in [6] pointed out to MWE categories that were mentioned by Sag et. al. in [14]. According to their classification there are two main types of MWEs:

- Lexicalized phrases: contains words that do not mean their own meaning. This category can be broken down into fixed, semi-fixed, and syntactically flexible expressions.

    1. Fixed expressions: This type has expressions such as: {*by and large, in short,* and *every which way*}
    2. Semi-fixed expressions: This type can be broken down into: non-decomposable idioms, compound nouns and proper nouns.
    3. Syntactically flexible expressions: verb-particle construction, decomposable idioms and light verbs. Verb-particle construction is a verb + one of more particles like: *brush up, break up*. Some forms can relocated the particles like in: *call him up, fight bravely on*. Decomposable idioms are syntactically flexible, examples include: *sweep under the rug, let the cat out of the bag.*. Finally, light verb constructions are like: *make a mistake, make a demo*.

- Institutionalized phrases: They are syntactically and semantically compositional but statistically idiosyncratic. A good example of it would be *traffic light* as this term can not be expressed with other words like: *traffic director* or *intersection regulator* because it has been conventionalized as it is. Thus, its idiosyncrasy is statistical rather than linguistic. Other examples of institutionalized phrases are: *telephone booth, fresh air.*

3. Related Work

In the context of MWE detection and integration some decent work has been introduced during the last four years. Ghoneim and Diab [6] described three methods to integrate MWEs into the Moses SMT system. Their work was an extension of Carpuat and Diab [4]. The study concentrated on how the integration methods are done, and focused less on how the MWE extraction process happens. MWEs were extracted from lexical databases, the English WordNet 3.0, and using named entity recognizers.

In the scope of automatically detecting MWEs, MWEtoolkit[4] [13] can be used in this task; because MWEtoolkit is a framework for language-independent MWE identification from corpora. Moreover, Attia et al. published an automatic technique to extract Arabic MWEs[2]. One of the coauthors of this work, an Arabic linguist, published a manually extracted Arabic MWEs on his personal website[5]. Despite the fact that this paper is concerned with detecting English MWEs in the training phase, there is a future work for this paper to detect MWEs in both sides of the parallel corpus and match them.

[4] it is a tool that aids in the automatic identification of multiword units such as idiomatic expressions (*kick the bucket*) and phrasal verbs (*take off, give up*) in large text bases, independently of the language

[5] http://www.attiaspace.com/