Artificial Intelligence 232 (2016) 54–75

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Topic-based term translation models for statistical machine translation

Deyi Xiong [a,*], Fandong Meng [b], Qun Liu [b,c]

[a] Soochow University, Suzhou, China
[b] Institute of Computing Technology, China
[c] School of Computing, Dublin City University, Ireland

ARTICLE INFO

Article history:
Received 24 August 2014
Received in revised form 9 December 2015
Accepted 14 December 2015
Available online 18 December 2015

Keywords:
Term
Term translation disambiguation
Term translation consistency
Term unithood
Statistical machine translation

ABSTRACT

Term translation is of great importance for machine translation. In this article, we investigate three issues of term translation in the context of statistical machine translation and propose three corresponding models: (a) a term translation disambiguation model which selects desirable translations for terms in the source language with domain information, (b) a term translation consistency model that encourages consistent translations for terms with a high strength of translation consistency throughout a document, and (c) a term unithood model that rewards translation hypotheses where source terms are translated into target strings as a whole unit. We integrate the three models into hierarchical phrase-based SMT and evaluate their effectiveness on NIST Chinese–English translation with large-scale training data. Experiment results show that all three models can achieve substantial improvements over the baseline. Our analyses also suggest that the proposed models are capable of improving term translation.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

A term is a linguistic expression that is used as the designation of a defined concept in a language (ISO 1087). The following sentences provide several term examples (in Italic).

Cambodia and Vietnam jointly hold *commodity exhibition*.
Indonesia reiterated its opposition to *foreign military presence*.
Native Mandarin speakers teach you *Chinese as foreign language*.

As shown in these examples, terms are compound words that are composed of nouns, adjectives and prepositions in special linguistic patterns.

As terms convey concepts of a text, appropriately translating terms is crucial when the text is translated from its original language to another language. The translations of terms are often affected by the domain in which terms are used and the context that surrounds terms [1]. In this article, we study domain-specific and context-sensitive term translation in the context of statistical machine translation (SMT).

* Corresponding author.
E-mail addresses: dyxiong@suda.edu.cn (D. Xiong), mengfandong@ict.ac.cn (F. Meng), liuqun@ict.ac.cn (Q. Liu).

http://dx.doi.org/10.1016/j.artint.2015.12.002
0004-3702/© 2015 Elsevier B.V. All rights reserved.

Table 1
Translation examples from the NIST MT02 Chinese-to-English test set. The underlined and underwaved words are source terms and their counterparts in baseline and reference translations, which highlight the three issues of term translation (ambiguity, consistency and unithood).

Eg. 1	Source	dan4 you2yu2 chang2gui1 sai4 yi3 lin2jin4 wei3sheng1, hua2sheng4dun4 qi2cai2 dui4 si4hu1 nan2yi3 yin1ci3 er2 chong1ji2 ji4 hou4 sai4
	Baseline	but because of conventional tournament is nearing an end, Washington Wizards team seems difficult to result after shocks quarter respectively
	Reference	However, as the regular season is approaching its end, it seems hard for Washington Wizards to impact the after season games as a result of this
Eg. 2	Source	yin4ni2 chong2shen1 fan3dui4 wai4guo2 jun1dui4 jin4zhu4... chong2shen1 fan3dui4 wai4guo2 jun1dui4 jin4zhu4 zhe4ge4 dao3guo2
	Baseline	Indonesia reiterates rejection of foreign military presence... reaffirming their opposition to foreign troops stationed in the island
	Reference	Indonesia Reiterated its Opposition to Foreign Military Presence... reiterated its opposition to foreign military presence in this island country

In order to achieve this goal, we focus on three issues of term translation: 1) ambiguity, 2) consistency and 3) unithood. First, term translation ambiguity is related to multiple translations of the same term in different domains. A source term may have different translations when it occurs in different domains. Second, term translation consistency is about consistent translations of terms that occur in the same document. Usually, it is undesirable to translate the same term in different ways as it occurs in different parts of a document. Finally, term unithood[1] concerns whether a multi-word term is still a unit after translation. Normally, a multi-word source term is translated as a whole unit into a contiguous target string.

Table 1 demonstrates the three issues of term translation with two Chinese-to-English translation examples. The first translation example (Eg. 1) visualizes two issues of term translation: ambiguity and unithood. In regard to the term translation ambiguity, the underlined source term "chang2gui1 sai4" can be translated into either "conventional tournament" or "regular season". The latter translation "regular season" is more widely used in the specific domain of NBA basketball games. Therefore given the domain of Eg. 1, "regular season" is a more appropriate translation for "chang2gui1 sai4" than "conventional tournament" that is chosen by the machine-generated baseline translation. As for the term unithood, the underwaved source term "ji4hou4 sai4" should be translated as a unit into target string "after season games". Unfortunately, the baseline translation violates the unithood constraint of this source term and translates it into an inconsecutive phrase that is interrupted by word "shocks".

The second translation example (Eg. 2) is related to term translation consistency. In this example, we display two sentences in the same text. The underlined source term "wai4guo2 jun1dui4 jin4zhu4" is not translated consistently in the baseline translations. It is translated as "foreign military presence" in the first sentence while "foreign troops stationed" in the second sentence (an undesirable translation).

In order to address these three issues of term translation, we propose a topic-based framework to model term translation for SMT. We capitalize on document-level topic information to disambiguate term translations in different documents and to maintain consistent translations for terms that occur in the same document. In particular, we propose the following three models.

- Term Translation Disambiguation Model: In this model, we condition the translations of source terms in different documents on the topic distributions of corresponding documents. In doing so, we enable the decoder to favor translation hypotheses with topic-specific term translations.
- Term Translation Consistency Model: We introduce a topic-dependent translation consistency metric for each source term to measure how consistently it is translated across documents in training data. With this metric, we encourage the same terms with a high strength of translation consistency that occur in different parts of a document to be translated in a consistent fashion.
- Term Unithood Model: We explore rich contextual information in the term unithood model to calculate how likely a source term should remain contiguous after translation. We use this unithood model to reward translation hypotheses where multi-word terms are translated as a whole unit.

A bilingual term bank is required to build these three models. We construct this term bank from our bilingual training data via automatic term extraction methods. We use a hierarchical phrase-based SMT system [3] to validate the effectiveness of the three term translation models. Large-scale experiment results show that they are all able to achieve substantial improvements of up to 0.88 BLEU points over the baseline. When simultaneously integrating the three models into SMT, we can gain a further improvement. The combination of the three models outperforms the baseline by up to 1.27 BLEU points.

The three term translation models have been first presented in our previous paper [4]. In this article, we make significant extensions to our previous work. First, for the purpose of completeness, we provide a background introduction of SMT and topic modeling, more details about bilingual term extraction, especially how we pair monolingual terms into bilingual terms

[1] Term unithood is defined as "*the degree of strength or stability of syntagmatic combinations and collocations*" by Kageura and Umino [2]. In this article we are interested in the unithood property of a target translation of a term.

Download English Version:

https://daneshyari.com/en/article/376782

Download Persian Version:

https://daneshyari.com/article/376782

Daneshyari.com