Baidu Translate: Research and Products

Zhongjun HE
Baidu Inc.
No. 10, Shangdi 10th Street, Beijing, 100085, China
hezhongjun@baidu.com

1 Overview

In this presentation, I would like to introduce the research and products of machine translation in Baidu. As the biggest Chinese search engine, Baidu has released its machine translation system in June, 2011. It now supports translations among 27 languages on multiple platforms, including PC, mobile devices, etc.

Hybrid translation approach is important for building an Internet translation system. As we know, the translation demands on the Internet come from various domains, including news wires, patents, poems, idioms, etc. It is difficult for a single translation system to achieve high accuracy on all domains. Therefore, hybrid translation is practically needed. Generally, we build a statistical machine translation (SMT) system, using the training corpora automatically crawled from the web. For the translation of idioms (e.g. *"有志者，事竟成,where there is a will, there is a way"*), hot words/expressions (e.g. *"一带一路, One Belt and One Road "*), example-based translation methods are used. To improve the translation of date (e.g. *"2012年7月6日, July 6, 2012"*), numbers (e.g. *"三千五百万, thirty-five million*), etc, rule-based methods are used as pre-process.

To improve translation quality for the resource-poor language pairs, we used pivot-based methods. Wu and Wang (2007) proposed the triangulation method that combines the source-pivot and the pivot-target phrase tables to induce a source-target phrase table. To fill up the data gap between the source-pivot and pivot-target corpora, Wu and Wang (2009) employed a hybrid method combining RBMT and SMT systems. We also proposed a method to use a Markov random walk to discover implicit relations between phrases in the source and target languages (Zhu et al., 2013), thus to improve the coverage of phrase pairs. We utilized the co-occurrence frequency of source-target phrase pairs to estimate phrase translation probabilities (Zhu et al., 2014).

On May 20th this year, we have launched a neural machine translation (NMT) system for Chinese-English translation. The system conducts end-to-end translation with a source language encoder and a target language decoder. Both the encoder and decoder are recurrent neural networks. The strength of NMT lies in that it can learn semantic and structural translation information by taking global contexts into account. We further integrated the SMT and NMT system to improve translation quality.

We also released off-line translation packs for NMT system on mobile devices, providing translation services in case that the Internet is unavailable. So far as we know, this is the first NMT system supporting off-line translation on mobile devices.

We also investigate the problem of learning a machine translation model that can simultaneously translate sentences from one source language to multiple target languages (Dong et al., 2015). Our solution is inspired by the recently proposed neural machine translation model which generalizes machine translation as a sequence learning problem. We train a unified neural machine translation model under the multi-task learning framework where the encoder is shared across different language pairs and each target language has a separate decoder. This model gets faster and better convergence for both resource-rich and resource-poor language pairs under the multi-task learning framework.

Based on the above techniques, we have released translation products for multiple platforms, including web translation on PC, APP on mobile devices, as well as free API for the third-party developers. Our system now support translations among 27 languages, not only including many frequently-used foreign languages, but also

Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra), pages 61–62,
Beijing, China, July 31, 2015.

including the traditional Chinese poem and Chinese dialects, for example, Cantonese. In order to make people communicate conveniently in foreign countries, the Baidu Translate APP supports speech-to-speech translation, object translation, instance full-screen translation, image translation, etc. Object translation enables users to identify objects and translate them into both Chinese and English. For the users who cannot speak and write foreign languages, the APP allows image as the input. For example, if you aim the cellphone camera at a menu written in a language you do not know, the translation will be displayed on the screen. Furthermore, rich information related to the food will also be displayed, including the materials, the taste, etc.

2 Outline

1. Introduction
    - Brief Introduction of Baidu MT
2. Hybrid Translation
    - SMT, EBMT and RBMT
    - Pivot-based Method for Resource-Poor Languages
3. Neural Machine Translation System
    - RNN Encoder-Decoder
    - Multi-task Learning
4. Products
    - Web
    - App
    - API

3 About the Speakers

Zhongjun He is a senior researcher in machine translation at Baidu. He received his Ph.D. in 2008 from Institute of Computing Technology, Chinese Academy of Sciences (ICT, CAS). He has more than ten years of research and development experiences on statistical machine translation.

References

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *To appear in ACL 2015*, Beijing, China, July.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 856–863.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th IJCNLP of the AFNLP*, pages 154–162.

Xiaoning Zhu, Zhongjun He, Hua Wu, Haifeng Wang, Conghui Zhu, and Tiejun Zhao. 2013. Improving pivot-based statistical machine translation using random walk. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, page 524534, Seattle, Washington, USA, October.

Xiaoning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang, and Tiejun Zhao. 2014. Improving pivot-based statistical machine translation by pivoting the co-occurrence count of phrase pairs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1665–1675, Doha, Qatar, October.