Machine Learning Approaches for Dealing with Limited Bilingual Data in Statistical Machine Translation

Gholamreza Haffari

1 Tutorial Description

Statistical machine translation (SMT) systems have made great strides in translation quality. However, high quality translation output is dependent on the availability of massive amounts of parallel text in the source and target language. There are a large number of languages that are considered "low-density", either because the population speaking the language is not very large, or even if millions of people speak the language, insufficient online resources are available in that language. This tutorial covers machine learning approaches for dealing with such situations in statistical machine translation where the amount of available bilingual data is limited. A statistical translation system can be *improved* and/or *adapted* by incorporating new training data in the form of parallel text.

The problem of learning from insufficient labeled training data has been dealt with in machine learning community under two general frameworks: (i) Semi-supervised Learning, and (ii) Active Learning. The goal of semi-supervised learning is to take advantage of abundant and cheap *unlabeled data*, together with labeled data, to build a high quality mapping from examples (the input space) to labels (the output space). On the other hand, the goal of active learning is to reduce the amount of labeled data required to learn a high quality mapping by querying the *user* to label the most informative examples so that the mapping is learnt with lesser number of examples.

The complex nature of machine translation task poses severe challenges to most of the algorithms developed in machine learning community for these two learning scenarios. This tutorial covers the subset of those methods and techniques from machine learning which have been successfully employed for statistical machine translation, together with the challenges they have faced. The tutorial covers a very recent area of research in SMT that has resulted in many publications over the last few years. As such, this topic has not been covered in a tutorial-level introduction elsewhere, and this would be the first such tutorial on the topic.

The tutorial is structured as follows. First we discuss the problem with respect to the availability of extra monolingual/multilingual data, the goals, and whether to focus on improving word alignment to improve SMT when the bilingual data is scarce. Then we mention the methods by classifying them into those belonging to semi-supervised learning and those belonging to active learning. For each of these learning scenarios, we cover (i) the background including the machine learning techniques used together with the task specific insights and observations, (ii) methods developed for single language-pair, and (iii) methods developed for multiple language-pairs. We carefully examine experimental results in two major conditions: (i) improving the SMT quality where the available bilingual data is scarce, and (ii) adapting to a new domain where enough bilingual data is available for one domain but scarce for the domain of interest. Finally, we point out some challenges in pushing further the state of the art in this line of research.

2 Timeline and Outline

- (20 mins) Introduction
  - Single Language-Pair vs Multiple Language-Pair
  - Goals: Improving Quality vs Adapting to a New Domain
  - Improving Word Alignment to Improve SMT? [10]
- (70 mins) Semi-supervised Learning (SSL) for SMT
  - Background
    - Inductive Learning vs Transductive Learning [4]
    - Expectation-Maximization, Self-Training, Co-Training [1, 13, 2]
  - Techniques for Single Language-Pair
    - SSL for Word Alignment [9]
    - Self-Training for SMT [18, 20, 19]
    - Paraphrasing [5]
  - Techniques for Multiple Language-Pairs
    - Co-Training/Coaching for SMT [3]
    - Triangulation [7]
- (10 mins) Break
- (70 mins) Active Learning with Selective Sampling for SMT
  - Background
    - What is Active Learning with Selective Sampling? [6]
    - Insights: Exploitation vs Exploration
    - Exploration: Expanding the Set of Lexicons/Phrases
    - Exploitation: Accurate Estimation of Translation Probabilities
    - Methods Dependent to/Independent from the Target Language
  - Techniques for Single Language-Pair [11]
    - Query by Committee [15]
    - Similarity to Bilingual Data
    - Decoder's Confidence [14]
    - Reconstruction of the Source
    - Hierarchical Adaptive Sampling [8]
    - Phrases/n-gram based Methods (for phrase-based SMT models) [12, 16]
    - Ensemble of Simple Models
  - Techniques for Multiple Language-Pairs [12]
    - A Unified Framework with Self-Training/Co-Training
    - Query by Committee
    - Combined Rankings [17]
- (10 mins) Concluding Remarks
- Total of 180 mins (3 hrs)

3 The Speaker

Gholamreza Haffari, PhD Candidate, School of Computing Science, Simon Fraser University
Phone: +1 (604) 636-3317
Email: ghaffar1@cs.sfu.ca
Web: http://www.cs.sfu.ca/∼ghaffar1/personal

Area of expertise:
Reza's research has been focused on machine learning algorithms applied to the study of natural language. He is especially interested in algorithms that combine labeled and unlabeled data, and learn new information with weak supervision for complex NLP problems: those involving *structured* output/*latent* variables such as Machine Translation.

On the topic of semi-supervised learning and active learning methods applied to NLP, he has published 1 book chapter (*Learning Machine Translation*), 1 journal article (*Machine Translation Journal*), and 5 conference papers (ICML, UAI, ACL, COLING, NAACL). His PhD dissertation, which is supervised by Prof. Anoop Sarkar, is entitled *Dealing with limited Training Data in Statistical Natural Language Processing.* Together with Anoop Sarkar, he has delivered a *tutorial on semi-supervised learning for natural language processing* in NAACL 2006. A full list of his papers is available at the web page listed above.

(please ignore the references included below, they are included only as a possible source of additional information to the proposal reviewer about our Schedule given above)

References

[1] Steven Abney. Understanding the yarowsky algorithm. *Computational Linguistics*, 2004.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[3] Chris Callison-Burch. Co-training for statistical machine translation. In *Master's Thesis, University of Edinburgh*, 2002.

[4] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[5] Philipp Koehn Chris Callison-Burch and Miles Osborne. Improved statistical machine translation using paraphrases. In *NAACL*, 2006.

[6] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. In *Machine Learning Journal*, 1994.

[7] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ACL*, 2007.

[8] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *ICML*, 2008.

[9] Alexander Fraser and Daniel Marcu. Semi-supervised training for statistical word alignment. In *ACL*, 2006.

[10] Kuzman Ganchev, Joao Graca, and Ben Taskar. Better alignments = better translations? In *ACL*, 2008.

[11] Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. Active learning for statistical phrase-based machine translation. In *NAACL*, 2009.

[12] Gholamreza Haffari and Anoop Sarkar. Active learning for multilingual statistical machine translation. In *submission*.

[13] Gholamreza Haffari and Anoop Sarkar. Analysis of semi-supervised learning with the yarowsky algorithm. In *UAI*, 2007.

[14] R.S.M. Kato and E. Barnard. Statistical translation with scarce resources: a south african case study. *SAIEE Africa Research Journal*, 98(4):136–140, December 2007.

[15] David Kauchak. Contribution to research on machine translation. In *PhD Thesis, UCSD*, 2006.

[16] Behrang Mohit and Rebecca Hwa. Localization of difficult-to-translate phrases. In *The Proceedings of the 2nd ACL Workshop on Statistical Machine Translations*, 2007.

[17] Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In *ACL*, 2008.

[18] Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Transductive learning for statistical machine translation. In *ACL*, 2007.

[19] Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Semi-supervised learning for machine translation. In Marc Dymetman Cyril Goutte, Nicola Cancedda and George Foster, editors, *Learning Machine Translation*. MIT Press, 2008.

[20] Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Semi-supervised model adaptation for statistical machine translation. *Machine Translation Journal*, 2008.