“Pivot Machine Translation between Statistical and Black box systems”

Final report

Antonio Toral

School of Computing, Dublin City University

Dubin, Ireland

atoral@computing.dcu.ie

31/03/2014

Introduction

This document reports on the project “Pivot Machine Translation between Statistical and Black box systems”, funded by the European Association for Machine Translation under its 2011 Call for Proposals program.

The project initially run from January to December of 2012 and was then granted a no-cost extension. This final report builds upon the mid-progress report, which covered the first phase (activities carried out in the first 5 months of the project).

Progress and Milestones Achieved in the Project

First Phase

Pivot Machine Translation (MT) refers to the use of an intermediate language, pivot language (PL), to translate from the source (SL) to the target language (TL). Much of the research carried out regards the scenario where both systems are statistical and it is assumed that the user has access not only to the output of the systems but also to internal data structures.

In this project we have developed a novel methodology for pivot-based MT which broadens the applicability of this technique, as it does not require internal access to the second MT system in the sequential pipeline (the one that translates from PL to TL), i.e. the second system is treated as a black box. Therefore, our novel approach allows to use systems other than statistical, such as rule-based.

We evaluated this approach on two language pairs (Italian to Catalan pivoting through Spanish and English to Macedonian pivoting through Bulgarian), obtaining 11-13% relative improvement in terms of BLEU over the baseline. A paper describing the algorithm and the experiment was presented at the EAMT 2012 conference.

Second Phase

In the second phase of the project we worked with the synthetic approach to pivoting. In this method, a SL–TL corpus is obtained using the SL–PL or the PL–TL corpora. One way to do this is to translate the PL sentences in the SL–PL corpus into TL with the PL–TL system. Another possibility is to translate the PL sentences in the PL–TL corpus into SL with the SL–PL system. Obviously, both methods could be applied and the two resulting synthetic corpora be merged into a single SL–TL corpus.

In the approach, typically, all the synthetic data generated is then used to train a statistical MT system from SL to TL. Our proposal was to use quality estimation to select a subset of the synthetic data of high translation quality. In this respect, our work can be considered as a first step towards the generation of reliable synthetic parallel data for under-resourced languages. Our methodology is as follows:

1. We first collect small amounts of aligned parallel data for the PL—TL language pair in order to build a quality estimation system.
2. We then translate the PL side of a SL—PL parallel corpus to TL with a PL—TL MT system.
3. We use the quality estimation system built in (1) to rank the translations obtained in (2) according to their estimated quality.
4. We build a SL—TL statistical MT system using a subset of the synthetic data ranked in (3), e.g. the first *n* sentence pairs.

We run experiments for the English—Croatian language pair with Slovene as the PL. We used a rule-based MT system (PL to TL) to translate the PL side of a SL—PL parallel corpus (Europarl) to the TL. We showed significant improvement in terms of automatic metrics obtained on two test sets using our approach compared to a random selection of synthetic parallel data. A paper describing the methodology and the experiment was presented at the LREC 2014 conference.

Publications

Raphael Rubino, Antonio Toral, Nikola Ljubešić and Gema Ramírez-Sánchez. 2014. **Quality Estimation for Synthetic Parallel Data Generation**. *In Proceedings of the 9th Language Resources and Evaluation Conference*. Reykjavik (Iceland). May 2014.

Antonio Toral. **Pivot-based Machine Translation between Statistical and Black Box systems**. *In Proceedings of the 15th Annual Conference of the European Association for Machine Translation*. Trento (Italy). May 2012.