An Enhanced Evaluation Metric for Statistical Machine Translation

Kirk Trombley John Morgan

August 5, 2014

Abstract

A popular evaluation metric for Statistical Machine Translation is enhanced with lexical meaning knowledge. The enhancement only requires a large monolingual corpus in the target language.

1 Introduction

Many Army operations involve a need to translate English text into a foreign language. For example in the case where Army special forces train foreign army units the texts used in these training operations should be available in the language of the foreign army.

Depending on its direction, translation can have two purposes for a country: assimilation, from a foreign language into the country's native language, or dissemenation, from the country's native language into a foreign language. Assimilation is highly valued in the Intelligence Community where analysts want to know what is being written in foreign texts. Dissemenation of English into low resourced foreign languages does not receive the attention given to assimilation. We are interested in using SMT for dissemenation purposes in Army training operations. Development of evaluation metrics is an aspect of dissemination that would benefit from more attention.

The traditional BLEU[**?**] metric uses exact string matching to score smt decoder output with reference transcriptions. This makes BLEU language-independent and thus is equally benefitial for assimilation and dissemenation purposes. BLEU is well known to have weaknesses[2]. Work has been done to remedy these weaknesses by incorporating llinguistic features into the evaluation metric. These enhancements make the metrics more accurate and correlate more with human judgements.

Meteor[1] is an evaluation metric that takes into consideration stem, synonym, and paraphrase matches between words inaddition to exact string matches. Evaluation metrics like meteor exist for resource-rich languages [1], but rare for

[1] Although see: `http://www.cs.cmu.edu/~mdenkows/meteor-universal.html` for an interesting development that makes meteor available in other languages. Note that this tool requires a parallel corpus large enough to train a moses smt system.

low-resourced languages of interest to the Army. For dissemmenation purposes, We would like to have an evaluation metric similar to Meteor that does not depend on high powered nlp tools like parsers or pos taggers which are not readily available in most languages.

Meteor extends BLEU by incorporating semantic features into the evaluation of SMT output. We similarly enhance BLEU with semantic knowledge by replacing words in the reference transcriptions with words that are close in meaning.

The word2vec tool is used to build a vector space word distribution model of the target language. Then a distance function is used to generate a neighborhood of words that are close to the words in the reference transcriptions. Word2vec only has one language-dependent requirement:a large monolingual corpus of text. It does not require a parser, a stemmer, or a part of speech tagger. The representations of languages are learned using the distributed Skip-gram or Continuous Bag-of-Words (CBOW) models recently proposed by . These models learn word representations using a neural network architecture that predicts the neighbors of a word.

The traditional BLEU metric measures the overlap of n-grams from the decoder output with n-grams from the reference transcription sentences. Credit is given to the decoder output n-gram only when there is an exact string to string match. If the reference sentence was:

The boy went to the store

But the decoder output:

The boy went to the supermarket

BLEU would assign no credit to the 1-gram supermarket. Although supermarket does not deserve the full creditof 1 given to store by BLEU, it seems harsh to give it no credit at all. Word2vec might place supermarket in a neighborhood of store where the two words have vectors that have a distance of 0.2 from each other. Our algorithm would assign 0.8 ($1 - 0.2$) instead of 0 to the word supermarket. We do this until we find a match for a small number ($k = 3$) of words in a neighborhood of the 1-gram in the reference 1-gram. For n-grams with $n > 1$ the algorithm is slightly more complicated. We first discard decoder output n-grams that differ by more than a single word. Then we consider the n-grams that result by replacing the word that differs with words close to it as above. Again we do this up to $k = 3$ times for the top k words that appear in the neighborhood of the differing word in the reference transcription.

References

[1] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[2] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report, September 17 2001.