Using a Paraphraser to Improve Machine Translation Evaluation

Reporter: Weigang LI

2004-4-18

Basic Information

- Author: Andrew Michael Finch
  Yasuhiro Akiba
  Eiichiro Sumita.
- Source: IJCNLP-2004

Content

- Paraphrasing system to improve automatic MT evaluation by augmenting the reference set with additional paraphrases of the human-produced references
- A data-oriented paraphraser
- Spearman rank correlation
- MT evaluation method

Introduction

- A single source sentence can be translated into many different ways
- Preparing references by hand is an expensive process
- This paper examines the usefulness of applying an automatic paraphraser to augment the references for automatic MT evaluation

A Data-oriented Paraphraser (DOPP)

- Based on the principles of data-oriented translation
- Directly translate sentences from one language into the same language, without going through an intermediate language
- The translation knowledge corpus is constructed by the sub-pattern

Data-oriented Translation (DOT)

- DOT is derived from data-oriented parsing(DOP)
  - DOP is a memory-based approach to parsing
  - Fragments (or subtrees) of parse trees are extracted from a training corpus of parsed sentences
  - These fragments are used as a grammar to parse unseen sentences
- Two trees are constructed at the same time
- The fragments contain links (between semantically equivalent nodes)-which are aligned automatically

Paraphrase Derivation

- Using a chart parser
  - Parse the input sentence and get the chart representation of the source sentence
  - Using the corresponding relation between the source and target sentences to get the chart representation of target sentence
  - Multi-candidate references are derived
  - Search the best reference
- Disambiguation
  - The most probable paraphrase is not sufficient
  - Using Monte-Carlo sampling to estimate the paraphrase probabilities

Experiments Overview

- A DOPP was trained on a corpus of English sentence pairs which are PP of each other
- Reference sets
  - 1-16 human-produced references
  - 1-100 their most probable paraphrases
- The output from nine different MT systems is then evaluated using each of scoring systems
- Scored sentences were analyzed for spearman rank correlation with judges

The aim

- To determine whether the paraphrases increased the correlation with the human ranks
- An increase in correlation indicates that the automatic evaluation system is more similar to a human in ranking the MT output

Spearman Rank Correlation

- 斯皮尔曼等级相关-Spearman rank correlation
- 适用于两个变项都是次序变项的数据时，通常，使用在计算两组等级之间一致的程度，如两个评分者评*N*件作品，或同一个人先后两次评*N*件作品等
- 此处计算自动MT评价和人工评价的相关度

Data

- Training data
  - ATR paraphrase corpus (about 50,000 sentences)
- Test data
  - 345 English sentences which are translated by nine different J-E MT systems
  - Scored by nine native English speaker
  - The median grade from nine grades is assigned by the human judges
- Reference data
  - 16 human-produced reference translations
  - These 16 sentences are paraphrased

Automatic Scoring Methods

- BLEU
- Multi-reference word error rate (mWER)
- Corrected Spearman Rank Correlation
  - Absolute values of the scores
  - Ordering->position->SRC

MT systems

- SMT
- TDMT
- EBMT
- Nine MT systems

Results

- The performance was enhanced until the number of paraphrases reached an optimal value
- 1 human-produced reference and 10 paraphrases is roughly equal the effect of 4-human-produced references