Sentence-Level Quality Estimation for MT System Combination

Tsuyoshi Okita, Raphaël Rubino, Josef van Genabith
Dublin City University

CNGL Centre for Next Generation Localisation
NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD Dublin University College Dublin
University of Limerick
Trinity College Dublin

Overview

Introduction

- Our approach: sentence-level Quality Estimation (QE) for system combination
- Two main steps
  1. Estimate sentence-level quality score for the 4 MT systems
  2. Pick the best sentence and use it as a backbone for system combination
- Two systems submitted
  1. Sentence-level system combination based on QE
  2. Confusion network based system combination

Introduction – Quality Estimation for MT

- How to estimate the translation quality when no references are available?
- First work at the word and sentence levels [**?**, **?**]
- More recently, WMT12 shared task on QE [**?**]
- State-of-the-art approach based on feature extraction and machine learning.

Overview

Sentence Level QE

- The aim is to estimate sentence-level TER scores for the 4 systems outputs
- Train set used to build regression model, TER estimation on test set
- Different features are extracted from the source and target sentence pairs
- We do not use provided annotations
- SVM used: ϵ-SVR with a Radial Basis Function kernel

Features Extraction – Adequacy and fluency

From the source and target sentences, we extract

- Surface features: sentence length, words length, punctuation, etc.
- Source and target surface features ratio
- Language model features: *n*-gram log-probability, perplexity
- Edit rate between the 4 MT outputs
- Two feature sets are built
  - **R1** constrained to provided data, contains target LM features and edit rates
  - **R2** unconstrained, contains all the features

Features Extraction – MT Output Edit Rate

For each MT system output, measure the edit rate with the three other systems' output.

System 1	Surprisingly, has checked that the new councillors almost do not comprise these known concepts.
System 2	Surprisingly, it has been proved that the new town councilors do almost not understand those known concepts.

Ins	**Del**	**Sub**	**Shft**	**WdSh**	**NumEr**	**NumWd**	**TER**
3	0	4	1	1	8.0	14.0	57.1

TER Estimation

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|ref_i - pred_i| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ref_i - pred_i)^2}$$

	system 1		system 2		system 3		system 4	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
R1	0.19	0.26	0.21	0.29	0.17	0.24	0.18	0.25
R2	0.20	0.26	0.21	0.29	0.21	0.28	0.20	0.26

Table: Error scores of the QE model when predicting TER scores at the sentence level on the test set for the four MT systems.

Overview

Standard System Combination Procedures (1)

- Procedures: For given set of MT outputs,
  1. (Standard approach) Choose backbone by a MBR decoder from MT outputs $\mathcal{E}$.

$$\begin{aligned}\hat{E}_{best}^{MBR} &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\ &= \text{argmin}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \quad (1) \\ &= \text{argmax}_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} BLEU_E(E') P(E|F) \quad (2)\end{aligned}$$

  2. Monolingual word alignment between the backbone and translation outputs in a pairwise manner (This becomes a confusion network).
  3. Run the (monotonic) consensus decoding algorithm to choose the best path in the confusion network.

Standard System Combination Procedures (2)

segment 3	
Input 1	they are normally on a week .
Input 2	these are normally made in a week .
Input 3	este himself go normally in a week .
Input 4	these do usually in a week .
Input 5	they are normally in one week .
Backbone(2)	these are normally made in a week .

Backbone(2)	these	are	normally	made	in	a	week .
hyp(1)	they$_S$	are	normally	*****$_D$	on$_S$	a	week .
hyp(3)	este$_S$	himself$_S$	go$_S$	normally$_S$	in	a	week .
hyp(4)	these	*****$_D$	do$_S$	usually$_S$	in	a	week .
hyp(5)	they$_S$	are	normally	*****$_D$	in	one$_S$	week .
Output	these	are	normally	*****	in	a	week .

Our Procedures of System Combination

- Procedures: For given set of MT outputs,
  1. Select backbone by QE.

$$\hat{E}_{best}^{QE} = \text{argmax}_{E' \in \mathcal{E}} QE(E')$$

  2. Monolingual word alignment between the backbone and translation outputs in a pairwise manner (This becomes a confusion network).
  3. Run the (monotonic) consensus decoding algorithm to choose the best path in the confusion network.

Overview

Results

	NIST	BLEU	METEOR	WER	PER
s1	6.50	0.225	0.5459	64.24	49.98
s2	6.93	0.250	**0.5853**	62.92	48.01
s3	7.40	0.245	0.5545	58.07	44.02
s4	7.21	0.253	0.5597	59.39	44.52
System combination without QE (standard)					
sys	**7.68**	0.260	0.5644	56.24	41.54
System combination with QE (1st algorithm)					
R1	**7.68**	**0.262**	0.5643	**56.00**	**41.52**
R2	7.51	0.260	0.5661	58.27	43.10
Backbone Performance (2nd Algorithm)					
R1	7.46	0.250	0.5536	57.68	43.38
R2	7.48	0.253	0.5582	57.76	43.28

Discussion (1)

	NIST	BLEU	METEOR	WER	PER
avg. TER	7.62	0.264	0.5653	56.40	41.61
s2 backbone	7.64	0.265	0.5607	56.01	42.01

Table: This table shows the performance when the backbone was selected by average TER and by one of the good backbone.

Discussion (2)

	System Combination TER Degradation (Case A)
source	"Me voy a tener que apuntar a un curso de idiomas", bromea.
QE	'I am going to have to point to a language course "joke.
comb	I am going to have to point to a of course ", kids.
ref	"I'll have to get myself a language course," he quips.
	System Combination TER Improvement (Case B)
source	Sorprendentemente, se ha comprobado que los nuevos concejales casi no comprenden esos conocidos conceptos.
QE	Surprisingly, it appears that the new councillors almost no known understand these concepts.
comb	Surprisingly, it appears that the new councillors almost do known understand these concepts.
ref	Surprisingly, it turned out that the new council members do not understand the well-known concepts.

Overview

Conclusions

- We presents two methods to use QE method.
  - for backbone selection in system combination.(1st algorithm)
  - for selection of sentence among translation outputs. (2nd algorithm)
- 1st algorithm
  - improvement of 0.89 BLEU points absolute compared to the best single system
  - 0.20 BLEU points absolute compared to the standard system combination strategy
- 2nd algorithm: lost of 0.30 BLEU points absolute compared to the best single system.
- At first sight, our strategy seemed to work quite well.

Acknowledgement

Thank you for your attention.

- This research is supported by the the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME project (Grant agreement No. 249119).
- This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.