Spoken Language Translation Graphs Re-decoding using Automatic Quality Assessment

Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Luong, Ngoc Tien Le

Spoken Language Translation Graphs Re-decoding using Automatic Quality Assessment

Laurent Besacier[1], Benjamin Lecouteux[1], Ngoc Quang Luong[1], Ngoc Tien Le[1]

[1]Grenoble Informatics Laboratory (LIG), Univ. Grenoble Alpes, Grenoble, France

contact: laurent.besacier@imag.fr

Abstract

This paper investigates how automatic quality assessment of spoken language translation (SLT) can help re-decoding SLT output graphs and improving the overall speech translation performance.
Using robust word confidence measures (from both ASR and MT) to re-decode the SLT graph leads to a significant BLEU improvement (more than 2 points) compared to our SLT baseline (French-English task).

Introduction

- **Context**
  - Automatic quality assessment of spoken language translation (SLT)
  - Also named confidence estimation (CE)
  - Pointing out correct parts and errors in a speech translated output
- **Useful for**
  - Interactive speech to speech translation
  - Computer-assisted translation (from speech or text)
- **Claim**
  - An accurate CE can also help to improve SLT itself through **search graph re-decoding**

Formalisation

x_f source signal, $f = (f_1, f_2, ..., f_M)$ transcription of x_f.
$\hat{e} = (e_1, e_2, ..., e_N)$ translation of f and $\hat{e} = \underset{e}{\text{argmax}}\{p(e/x_f, f)\}$

Word Confidence Estimation (**WCE**) can be seen as finding sequence q where $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$
$\hat{e} = \underset{e}{\text{argmax}} \sum_q p(e, q/x_f, f)$ $\hat{e} = \underset{e}{\text{argmax}} \sum_q p(q/x_f, f, e) * p(e/x_f, f)$

$$\hat{e} \approx \underset{e}{\text{argmax}}\{\underset{q}{\text{max}}\{\boldsymbol{p(q/x_f, f, e) * p(e/x_f, f)}\}\}$$

$p(q/x_f, f, e)$: WCE component
$p(e/x_f, f)$: SLT component

SLT Search Graph (SG) Re-decoding

For SLT N-best hypotheses $e^N = \{e^1, e^2, ..., e^N\}$, each *j-th* word in the hypothesis e^i, denoted by e_{ij}, has a quality label, q_{ij}.
For all hypothesis H_k in the SG, the new transition cost is defined by:

$$transition'(H_k) = transition(H_k) + \begin{cases} reward(e_{ij}) & \text{if } q_{ij} = good \\ penalty(e_{ij}) & otherwise \end{cases} \quad (1)$$

with

$$penalty(e_{ij}) = -reward(e_{ij}) = \beta * \frac{score(H_k)}{\#words(H_k)} \quad (2)$$

Analysis of SLT Hypotheses

example 1	
f_{ref}	une démobilisation des employés peut déboucher sur une démoralisation **mortifère**
f_{hyp}	une démobilisation des employés peut déboucher sur une démoralisation **mort y faire**
e_{hyp} baseline	a **demobilisation employees** can lead to a **penalty demoralisation**
e_{hyp} with re-decoding	a **demobilisation of employees** can lead to a **demoralization death**
e_{ref}	**demobilization of employees** can lead to a **deadly demoralization**
example 2	
f_{ref}	celui-ci a indiqué que l'intervention **s'**était parfaitement bien **déroulée** et que les examens post-**opératoires** étaient normaux
f_{hyp}	celui-ci a indiqué que l' intervention **c'**était parfaitement bien **déroulés** , et que les examens post **opératoire** étaient normaux .
e_{hyp} baseline	**it** has indicated that the speech **that was well** conducted , and that the tests were **normal post route**
e_{hyp} with re-decoding	**he** indicated that the intervention **is very well done** , and that the tests **after operating were normal**
e_{ref}	**he** indicated that the operation **went perfectly well** and the **post-operative tests were normal**
example 3	
f_{ref}	general motors repousse jusqu'en janvier le plan pour **opel**
f_{hyp}	general motors repousse jusqu' en janvier le plan pour **open**
e_{hyp} baseline	general motors postponed until january **the plan to open**
e_{hyp} with re-decoding	general motors puts until january **terms to open**
e_{ref}	general motors postponed until january **the plan for opel**

Table : Exemples of French SLT output with and w/o re-decoding

Building an Efficient WCE System

$\hat{q} = \underset{q}{\text{argmax}}\{p(q/x_f, f, e)\}$
need training data with quadruplet (x_f, f, e, q) available,
so instead we compute: $\hat{\boldsymbol{q}} = \underset{\boldsymbol{q}}{\textbf{argmax}}\{\boldsymbol{p_{ASR}(q/x_f, f)^{\alpha} * p_{MT}(q/e, f)^{1-\alpha}}\}$

- $p_{ASR}(q/x_f, f)$
  - system described in [1]
  - acoustic / graph / linguistic / lexical features
  - boosting classifier
- $p_{MT}(q/e, f)$
  - system described in [2]
  - multiple features and CRF classifier
  - using our open-source toolkit available online
    `http://github.com/besacier/WCE-LIG`

Experimental Setting

- **French ASR**
  - Kaldi toolkit [3]
  - HMM/SGMM / 3-grams
- **Test corpus**
  - 2643 French utterances
  - 5h of speech
  - Cross-validation for tuning / decoding
  - Quality labels $q_i \in \{good, bad\}$) obtained with TERp-A toolkit [4]
- **French-English SMT**
  - Moses toolkit [5]
  - 1.6M parallel sent.
  - 48M monolingual sent.
  - medium-size system
  - WMT shared task

task	ASR (WER)	MT (BLEU)	% good	% bad
MT	0%	52.8%	82.5%	17.5%
SLT	26.6%	30.6%	65.5%	34.5%

Table : Baseline MT and SLT performance on 2643 utt.

Re-decoding Results

- **2-pass (graph re-decoding)**
  - MT features only
  - ASR features only
  - joint MT+ASR features

system	baseline	redecoding	redecoding	redecoding
WCE feat.	**none**	**ASR** $p(q/x_f, f)$	**MT** $p(q/f, e)$	**SLT** $p(q/x_f, f, e)$
Perf.	30.60%	31.12%	31.89%	**32.82%**

Table : SLT perf. (BLEU) after 2d pass (2643 utt.)

References

[1] Laurent Besacier, Benjamin Lecouteux, Ngoc Quang Luong, Kaing Hour, and Marwa Hadjsalah,
"Word confidence estimation for speech translation,"
in *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA, December 2014.

[2] Ngoc Quang Luong, Laurent Besacier, and Benjamin Lecouteux,
"An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation,"
in *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatie, jun 2014.

[3] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely,
"The kaldi speech recognition toolkit,"
in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[4] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz,
"Terp system description,"
in *MetricsMATR workshop at AMTA*, 2008.

[5] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst,
"Moses: Open source toolkit for statistical machine translation,"
in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 177–180.