

Disentangling ASR and MT Errors in Speech Translation

Ngoc-Tien Le, Benjamin Lecouteux, Laurent Besacier

▶ **To cite this version:**

Ngoc-Tien Le, Benjamin Lecouteux, Laurent Besacier. Disentangling ASR and MT Errors in Speech Translation. MT Summit, 2017, Nagoya, Japan. hal-02094763

HAL Id: hal-02094763
https://hal.science/hal-02094763

Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Disentangling ASR and MT Errors in Speech Translation

Ngoc-Tien Le, Benjamin Lecouteux, Laurent Besacier

Grenoble Informatics Laboratory (LIG), Univ. Grenoble Alpes, Grenoble, France

contact: firstname.lastname@univ-grenoble-alpes.fr

Abstract

Investigating automatic detection of SLT errors that can be due to transcription (ASR) or to translation (MT) modules.
Using robust word confidence measures (from both ASR and MT) to disentangle ASR and MT errors in the speech translation output.

Introduction

- **Context**
  - Automatic quality assessment of spoken language translation (SLT), called 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 disentangle ASR and MT errors** in the speech translation output.

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 $\hat{e} = \underset{e}{\text{argmax}}\{p(e|x_f, f)\}$

A quality estimation (or error detection) component in speech translation solves the equation: $\hat{q} = \underset{q}{\text{argmax}}\{p_{SLT}(q|x_f, f, \hat{e})\}$

Word Confidence Estimation (**WCE**) can be seen as finding sequence q where $q = (q_1, q_2, ..., q_N)$ and $q_i \in \{good, bad\}$

Experimental Setting

- **French ASR**
  - KALDI toolkit
  - CD-DNN-HMM acoustic model
  - 3-gram LM
  - Two LMs: 62K (ASR1) and 95K (ASR2)
- **French-English SMT**
  - Moses toolkit
  - 1.6M parallel sent.
  - 48M monolingual sent.
  - medium-size system
  - WMT shared task
- *** Corpus**
  - 6693 French utterances (2643 dev + 4050 tst)
  - 16h52 of speech (5h51 dev + 11h01 tst)
  - Quality labels $q_i \in \{good, bad\}$) obtained with TERp-A toolkit

Task	*ASR (WER)*		*MT (BLEU)*		*% G (good))*		*% B (bad)*	
	dev set	*tst* set	*dev* set	*tst* set	*dev* set	*tst* set	*dev* set	*tst* set
MT			49.13%	57.87%	76.93%	81.58%	23.07%	18.42%
SLT (ASR1)	21.86%	17.37%	26.73%	36.21%	62.03%	70.59%	37.97%	29.41%
SLT (ASR2)	16.90%	12.50%	28.89%	38.97%	63.87%	72.61%	36.13%	27.39%

Table: ASR, MT and SLT performances on our *dev* set and *tst* set.

Disentangling: Word Alignments between MT and SLT

Motivation: How many erroneous words - in the SLT output - is *a source word* aligned to?

where $\hat{e}_{slt} = (e_1, e_2, , e_n)$, $\hat{e}_{mt} = (e'_1, e'_2, , e'_m)$, $L = (l_1, l_2, , l_n)$: set of word alignments
$e_{hyp_{slt}} \leftrightarrow e_{hyp_{mt}}$ if existing word alignment $e_{k_j} \leftrightarrow e'_{k_i}$; (e_{k_j}, e'_{k_i}) = *False*, otherwise.

list_labels_result ← *empty_list*
for each sentence $e_k \in \hat{e}_{slt}$ **do**
  list_labels_sent ← *empty_list*
  for j ← 1 to *NumberOfWords*(e_k) **do**
    if *label*(e_{k_j}) = 'G' **then**
      add 'G' to *list_labels_sent*
    else if Existed Word Alignment (e_{k_j}, e'_{k_i}) and *label*(e'_{k_i})='B' **then**
      add '*B_MT*' to *list_labels_sent*
    else
      add '*B_ASR*' to *list_labels_sent*
    end if
  end for
  add *list_labels_sent* to *list_labels_result*
end for

Disentangling: Subtraction between SLT and MT Errors

Motivation: differences between SLT hypothesis ($e_{hyp_{slt}}$) and MT hypothesis ($e_{hyp_{mt}}$)

list_labels_result ← *empty_list*
for each sentence $e_k \in \hat{e}_{slt}$ **do**
  list_labels_sent ← *empty_list*
  for j ← 1 to *NumberOfWords*(e_k) **do**
    if *label*(e_{k_j}) = 'G' **then**
      add 'G' to *list_labels_sent*
    else if *NameOfWordAlignment*(l_{k_i}) is 'Insertion' OR 'Substitution' **then**
      add '*B_ASR*' to *list_labels_sent*
    else
      add '*B_MT*' to *list_labels_sent*
    end if
  end for
  add *list_labels_sent* to *list_labels_result*
end for

Example with 3-label Setting

$e_{hyp_{slt}}$	surgeons	in	los	angeles	it	is	said
$e_{hyp_{mt}}$	surgeons	in	los	angeles	**	have	said
edit op.	Exact	Exact	Exact	Exact	Insertion	Substitution	Exact

Table: Example of Edit Distance between SLT and MT.

f_{ref}	les chirurgiens	de	los	angeles	ont		dit
f_{hyp}	les chirurgiens	de	los	angeles	on		dit
labels ASR	G G	G	G	G	B		G
$e_{hyp_{mt}}$	surgeons	in	los	angeles		have	said
labels MT	G	B	G	G		B	G
$e_{hyp_{slt}}$	surgeons	in	los	angeles	it	is	said
labels SLT (2-label)	G	B	G	G	B	B	G
labels SLT (*Method 1*)	G	B_MT	G	G	B_ASR	B_MT	G
labels SLT (*Method 2*)	G	B_MT	G	G	B_ASR	B_ASR	G
e_{ref}	the surgeons	of	los	angeles			said

Table: Example of Quintuplet with 2-label and 3-label.

Statistics with 3-label Setting on the Whole Corpus

Task - *ASR1*	*dev* set			*tst* set		
	%G	*%B_ASR*	*%B_MT*	%G	%B_ASR	%B_MT
label/m1:Method 1	*62.03*	*19.09*	*18.89*	70.59	14.50	14.91
label/m2:Method 2	*62.03*	*22.49*	*15.49*	70.59	16.62	12.79
label/same(m1, m2)	*62.03*	*18.09*	*14.49*	70.59	13.58	11.88
label/diff(m1, m2)	*0*	*1.00*	*4.40*	0	0.92	3.03
Task - *ASR2*	***dev* set**			***tst* set**		
	%G	*%B_ASR*	*%B_MT*	%G	%B_ASR	%B_MT
label/m1:Method 1	*63.87*	*16.89*	*19.23*	72.61	11.92	15.47
label/m2:Method 2	*63.87*	*19.78*	*16.34*	72.61	13.58	13.81
label/same(m1, m2)	*63.87*	*16.05*	*15.50*	72.61	11.12	13.01
label/diff(m1, m2)	*0*	*0.84*	*3.73*	0	0.80	2.46

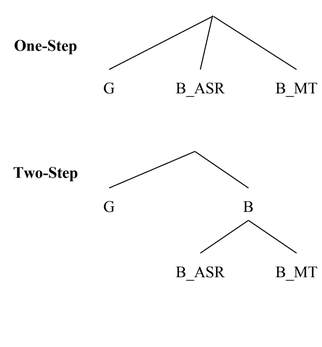
Table: Statistics with 3-label setting for *ASR1* and *ASR2*.

Experiments on 3-class Error Detection

2-class Full Corpus		3-class Intersection Corpus (m1, m2)				
			One-Step		Two-Step	
ASR1	*ASR2*		*ASR1*	*ASR2*	*ASR1*	*ASR2*
F_G 81.79	83.17	F_G	85.00	85.00	84.00	85.00
F_B 48.00	45.17	F_{B_ASR}	44.00	42.00	44.00	42.00
		F_{B_MT}	14.00	15.00	16.00	17.00
F_{avg} 64.90	64.17	F_{avg}	47.67	47.33	48.00	48.00

Table: Error Detection Performance (2-label vs 3-label) on SLT output for tst set (training is made on *dev* set).

*** One-Step vs Two-Step**



Conclusions

- Proposed **2 methods** for the non trivial label setting to disentangle ASR and MT errors in speech translation
- Recasting the binary error detection problem to **3-class labeling problem** (*good, asr-error, mt-error*)
- Using joint ASR and MT features, automatic detection of error types was evaluated and encouraging results were displayed on a French-English speech translation task
- Providing further support for building better informed speech translation systems, especially in interactive speech translation use cases