LIG English-French Spoken Language Translation System for IWSLT 2011

Benjamin Lecouteux, Laurent Besacier, Hervé Blanchon

LIG English-French Spoken Language Translation System for IWSLT 2011

Benjamin Lecouteux, Laurent Besacier and Hervé Blanchon

Abstract

This paper describes the system developed by the LIG laboratory for the 2011 IWSLT evaluation. We participated to the English-French MT and SLT tasks.
The development of a reference translation system (MT task), as well as an ASR output translation system (SLT task) are presented. We focus this year on the SLT task and on the use of multiple 1-best ASR outputs to improve overall translation quality. The main experiment presented here compares the performance of a SLT system where multiple ASR 1-best are combined before translation (source combination), with a SLT system where multiple ASR 1-best are translated, the system combination being conducted afterwards on the target side (target combination). The experimental results show that the second approach (target combination) overpasses the first one, when the performance is measured with BLEU.

Introduction

We focus on the SLT task and on the use of multiple 1-best ASR outputs to improve translation. Two different approaches are proposed:

→ source combination: multiple ASR 1-best are combined before translation

→ target combination: multiple ASR 1-best are translated, before applying system combination on the target side

LIG systems in 2010

→ TED Talks collection plus other parallel corpora distributed by the ACL 2010 WMT

→ Training of the translation models :

→ Europarl and News parallel corpora (1,767,780 sentences)

→ TED training corpus (total 47,652 sentences)
→ Language model training :
→ News-mono+TED-mono
→ 2010 News monolingual corpus in French (15,234,997 sentences)

→ TED dev set (934 sentences) : for tuning and evaluation purpose (Dev2010)
→ Do not reorder over punctuation during decoding
→ Phrase-table pruning

- **Phrase based translation model**
- Tokenised and lowercased corpora
- Kept punctuation for the LM and TM models
- Translation model: **Giza++** and **Moses**
- **Decoding: 8 scores** on feature functions
- **Language model** : 3-gram based, SRILM
- Weights tuning : **MERT** method on TED Dev2010, without punctuation

SLT task :

→ ASR output lowercased, tokenized and re-punctuated before translation

→ True re-punctuation system for French : LM trained on punctuated and uncased French data (Europarl+News+UN+Newsmono: 24M sentences)

→ Punctuation restored using hidden-ngram

→ SMT-based recaser presented earlier

Corpus	BLEU c+p Dev2010/ tst2010	BLEU c dev2010/ tst2010	BLEU x dev201/ tst2010
7. (6) + pre-/post-process described in 2.6	0.1670/ 0.2027	0.1606/ 0.1992	0.1709/ 0.2081
8. (7)+ tuning on ASR input (Dev2010)	0.1745/ 0.2087	0.1671/ 0.2046	0.1766/ 0.2133

MT and SLT LIG systems in 2011

System	BLEU c+p dev2010 /tst2010	BLEU c dev2010 / tst2010	BLEU x dev2010/ tst2010
1. LIG 2010 2010 bitexts	0.2408/0 .2758	0.2179/0 .2479	02311/ 0.2590
2. PT trained on TED2011 bitext only (no tuning)	0.2270/0 .2782	0.2044/0 .2508	0.2167/ 0.2611
3. PT trained on TED2011 bitext only (+tuning)	0.2411/0 .2781	0.2168/0 .2513	0.2296/ 0.2621
4. (1)+update LM using TED 2011 mono	0.2452/0 .2789	0.2207/0 .2516	0.2335/ 0.2623
5. Multiple PT - Either(1,4) - no tuning + updated LM	0.2397/0 .2898	0.2167/0 .2618	0.2293/ 0.2726
6. Multiple PT - Either(1,4) + tuning + updated LM	0.2524/0 .2896	0.2289/0 .2623	0.2420/ 0.2733

MT system :

→ Phrase-table trained on the TED 2011 bilingual data (107268 sentences) only with and without tuning (2,3)

→ Target LM : updated using the TED 2011 mono (111431 sentences) data (4)

→ retuning on dev2010, this approach improved the system by more than 1 point BLEU (5,6)

Source versus Target Combination

Source combination

→ classical ROVER weighted by the ASR WER :

*alpha*Sum(WordOcc) + (1-alpha)*Sum(Confidence(W))*

Where alpha=0.9 and confidence scores are empirically defined

System	WER%
0	17.1
1	18.2
2	17.4
3 (not used)	27.3
4	**15.3**

Target combination

→ 500-best translated outputs generated from each ASR source system

→ Moses option *distinct*

→ N-best associated with a set of 13 features (10 TM, 1 distance-based, 1 LM, 1 word penalty)
→ Combined in several steps
→ Score combination weights optimized on a dev corpus (BLEU at the sentence level)
→ N-best resorted using SRILM nbest-optimize

→ Once the optimized feature weights are computed independently for each ASR source :
→ N-best lists are turned into confusion networks (CN)
→ Features used to compute posteriors relatively to all the hypotheses in the N-best list
→ CN computed for each sentence and for each system
→ CN merged into a single one
→ ROVER is applied on the combined CN and generates a lowercased 1-best

→ When 3 systems are available, the target combination is better than the source combination

→ As more ASR systems (2, 3, 4) are added, the overall performance improves

→ source+target combination show a slight BLEU degradation

→ combination weights tuned on tst2010 data (bigger than dev2010)

→ **dev2010 considered as a validation test in this table**

Combination	BLEU c+p Dev2010/ tst2010	BLEU c dev2010/ tst2010	BLEU x dev2010/ tst2010
Sys 0 alone	0.1671/ 0.2012	0.1602/ 0.1957	0.1695/ 0.2039
Sys 1 alone	0.1608/ 0.1944	0.1534/ 0.1909	0.1622/ 0.1985
Sys 2 alone	0.1737/ 0.2027	0.1664/ 0.1975	0.1768/ 0.2072
Sys 4 alone	0.1770/ 0.2082	0.1709/ 0.2033	0.1811/ 0.2125
Target comb. (systems 42)	0.1772/ 0.2085	0.1710/ 0.2036	0.1812/ 0.2130
Source comb. (rover systems 420) done at LIG	0.1787/ 0.2139	0.1709/ 0.2099	0.1811/ 0.2191
Target comb. (systems 420)	0.1815/ 0.2136	0.1748/ 0.2087	0.1852/ 0.2178
Source comb. (rover systems 0213) provided by IWSLT orga. (cf tab 3)	0.1745/ 0.2087	0.1671/ 0.2046	0.1766/ 0.2133
Source comb. (rover systems 4021) done at LIG	0.1797/ 0.2159	0.1726/ 0.2115	0.1826/ 0.2209
Target comb. (systems 4021)	0.1841/ 0.2143	0.1782/ 0.2099	0.1889/ 0.2189
Source+Target comb. (systems 4021R)	0.1818/ 0.2166	0.1758/ 0.2120	0.1859/ 0.2215

Official results and Conclusion

System.	bleu(p+c)	bleu(x)
LIG_P (Tst2011) source+target comb. (4201R)	0,2485	0,2598
LIG_C1 (Tst2011) source comb. (4201)	0,2453	0,2561
LIG_PostEval (Tst2011) Target comb (4201)	0.2489	0.2599

→ English-French MT updated on the new data without radical changes

→ Several approaches to take advantage of multiple ASR system outputs

→ Results show that combining translation hypotheses on the target language side lead to better results than combining ASR 1-best on the source side, before translation (0.4 BLEU improvement observed)