ITI-UPV system description for IWSLT 2010

Germán Sanchis–Trilles, Guillem Gascó, Vicent Alabau, Jesús Andrés–Ferrer, Jesús González–Rubio
Martha–Alicia Rocha, Francisco Casacuberta, Jorge González, Joan–Andreu Sánchez
{gsanchis,ggasco,valabau,jandres,jegonzalez,mrocha,fcn,jgonzalez,jandreu}@dsic.upv.es

DSIC UNIVERSIDAD POLITECNICA DE VALENCIA PRHLT MIPRCV CONSOLIDER INGENIO 2010 Multimodal Interaction in Pattern Recognition and Computer Vision

Introduction

PRHLT group took part in the DIALOG and TALK tasks, emphasis on:

- DIALOG: Translation between syntactically different languages
  - Make use of syntactic and linguistic information by means of ITGs
- TALK: Efficient training of SMT models in resource rich scenarios
  - Include only appropriate training samples into the model

DIALOG task: ITGs and model combination

1. *I*nversion *T*ransduction *G*rammar-based decoding
   - Formalism that underlies the translation: Phrasal ITGs.
   - Chinese linguistic parsing enriches the reordering process.
2. Lattices for ASR error recovery
   - Training corpus preprocessed to resemble ASR input
   - Coupled ASR translation using Confusion Networks
3. Median string computation for hypothesis combination
   - Combine ITG-based and phrase-based translations
   - Benefit from high coverage of PB models and quality of ITGs
   - Consensus translation computed as the median string

TALK task: Selection in resource-rich scenarios

1. Subtitle segmentation recovery
   (a) Build independent sentences from subtitles.
   - Subtitles concatenated until end of line mark (".", "?","!").
   - Information about subtitles is kept by means of <wall/> tag

   (b) Translate input as a block, then recover subtitles
2. Probabilistic sentence selection
   - Two kind of corpora: *in-domain*, and *out-of-domain*
   - Estimate an in-domain probability model:
   $$p(\mathbf{e},\mathbf{f},|\mathbf{e}|,|\mathbf{f}|) = p(\mathbf{e},\mathbf{f}/|\mathbf{e}|,|\mathbf{f}|)p(|\mathbf{e}|,|\mathbf{f}|)$$
   - with $p(|\mathbf{e}|,|\mathbf{f}|)$ estimated by MLE, and
   - $p(\mathbf{e},\mathbf{f}/|\mathbf{e}|,|\mathbf{f}|)$ a log linear model with the features:
     - a direct and an inverse IBM model 4, and
     - both source and target, 5-gram language model
3. On-line sentence selection for infrequent n-grams recovery
   - Sentences with more infrequent n-grams are more informative
   - Dynamic update of the n-grams counts
   - Score each n-gram according to $\mathrm{s}(x) = \max\{0, t - N(x)\}$
     - → $N(x) =$ occurrence of n-gram x in test
     - → $t =$ minimum occurrence considered infrequent
   - Select sentences with maximum global $\mathrm{s}(x)$ score
4. Bayesian adaptation for model stabilization
   - MERT unstable if amount development data is small
   - Bayesian adaptation: weights Λ viewed as random variables.
   - After several approximations:
   $$p(\mathbf{e}|\mathbf{f};T,A) = \sum_{\Lambda_m \in MC(\Lambda_t)} \left(p(A|\Lambda;T)p(\mathbf{e}|\mathbf{f},\Lambda)\right)^{\frac{1}{\delta}} p(\Lambda|T)$$
   with δ a leveraging factor and $MC(\Lambda_t)$ a weight-sampling.

Description of the presented systems

- Baseline system
  - Built with Moses in standard setup
  - 5-gram LM with Kneser-Ney smoothing built with SRILM
  - Log-linear weights optimized with MERT
- DIALOG: Chinese–English system
  - Combination of various MT systems:
    - ∗ Phrase-based:
      - · All references in development used in training
      - · Only single reference for the language model
    - ∗ ITG-based: with linguistic information (Stanford parser)
    - ∗ ASR system (only for ASR output condition)
  - Normalized edit distance as dissimilarity measure
  - Combine the 20-best translations of each system
- TALK: English–French system
  - All experiments conducted with phrase-based models (Moses)
  - Training set split into training and internal development
  - Results here shown for official development set
  - All corpora provided were used for sentence selection
  - Baseline with only in-domain corpus: 23.2 BLEU

Results on probabilistic sentence selection. nK = thousands sentences added.

nK	BLEU
0	23.2
10	23.5
50	24.2
100	25.0
200	25.1
500	25.5

Results for infrequent n-gram recovery. $|S|$ = total number of training sentences.

nK	t	$\|S\|$	MERT	bayes
50	-	96.9K	24.2	24.7
	1	99.9K	23.7	24.9
	10	101.9K	24.1	25.2
100	-	146.9K	25.0	25.1
	1	149.8K	24.6	25.3
	10	156.9K	24.1	25.4

- Translation quality when using MERT shows to be unstable
- Bayesian adaptation proves to be able to stabilize the weights
- Probabilistic sentence selection provides improvements
- Infrequent n-gram recovery provides further improvements
- Final system: $nK = 500$ and $t = 10$, total of 645.6K sentences

Conclusions and future work

- DIALOG
  - Consensus translations combine strengths of PB models and ITGs
  - ASR: Confusion networks provided improvements over 1-best
  - Future work:
    - → Lexicalized Max. Entropy models for reordering in ITGs
    - → Research translation of lattices with confidence measures
- TALK
  - Intelligent data selection: better use of computational resources
  - Two BLEU points below best system, using only 3% of training
  - Bayesian adaptation can be applied for system stabilization
  - Future work:
    - → Compare selection vs random, optimize weights
    - → Research ratio probabilistic selection vs infrequent n-grams

Acknowledgments

Supported by the EC (FEDER/FSE), the Spanish MICINN and MITyC and the Generalitat Valenciana under projects MIPRCV, ALMPR, iTrans-2 and erudito.com.