PRHLT System Description for TALK Task

Germán Sanchis-Trilles
gsanchis@dsic.upv.es

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática • Universidad Politécnica de Valencia

2nd December 2010

Outline

- Overview of PRHLT submission
- About the TALK task
- Problems confronted
  - Probabilistic sentence selection
  - Infrequent n-grams recovery
  - Bayesian adaptation for model stabilization
- Experiments
- Conclusions and future work

PRHLT submission

- Authors: Guillem Gascó, Vicent Alabau, Jesús Andrés–Ferrer
  Jesús González–Rubio, Martha–Alicia Rocha, Germán Sanchis–Trilles
  Francisco Casacuberta, Jorge González, Joan–Andreu Sánchez

- Submitted runs for both DIALOG and TALK tasks

- For the DIALOG task, focus on:
  - ITGs for syntactically different languages
  - System combination between ITGs and PB
  - Lattice translation for ASR error recovery

About the TALK task

- Public speeches on several topics, English–French translation
- Sentences segmented at the subtitle level
- Small size of in-domain corpus ($\approx$ 45K sentence pairs)
- Large amount of out-of-domain corpora available ($\approx$25M sentence pairs)

Problems confronted

1. Subtitle translation rather than sentence translation

   ? Treat subtitles as sentences (not the same thing!)
   ! Re-build sentences, translate, recover subtitle segmentation

2. Large amount of out-of-domain corpora available

   ? Use all corpora available
   ⇒ In-domain information might be overwhelmed
   ⇒ Very expensive in computational terms
   ! Select sentences in a smart way
   ⇒ Probabilistic sentence selection (do not disturb in-domain distribution)
   ⇒ Infrequent n-gram recovery (increase informativeness of data)

3. Small amount of data turns models unstable

   ! Apply model stabilization techniques

Probabilistic sentence selection: Motivation

- Motivation: Corpora sizes grow faster than the computational resources needed
- Aim: Select most useful sentences from out-of-domain corpora
- Premises over the selected subcorpus:
  - Must not disturb excessively the probability distribution of the in-domain corpus
  - Must be informative
- Purpose:
  - Alleviate the high computational effort of using the whole corpus
  - Avoid overwhelming the in-domain data with too much out-of-domain corpora

Probabilistic Sentence Selection

- Approximate the probability distribution of the in-domain corpus by:

$$p(\mathbf{e}, \mathbf{f}, |\mathbf{e}|, |\mathbf{f}|) \approx p(\mathbf{e}, \mathbf{f}/|\mathbf{e}|, |\mathbf{f}|)p(|\mathbf{e}|, |\mathbf{f}|)$$

- Length distribution computed by maximum likelihood estimation
- $p(\mathbf{e}, \mathbf{f}/|\mathbf{e}|, |\mathbf{f}|) \approx \frac{1}{\mathcal{Z}(\mathbf{e}, \mathbf{f})} \exp(\sum_k \lambda_k h_k(\mathbf{e}, \mathbf{f}))$
- Estimate all models using just the in-domain corpora
- Sample from out-of-domain corpora without replacement

On-line Sentence Selection for Infrequent n-grams Recovery

- Alignment for n-grams that appear rarely in training cannot be estimated accurately

  $\Rightarrow$ Extreme case: out-of-vocabulary n-grams

- If such n-gram appears in test, it might not be translated accurately

  $\Rightarrow$ In real test set, 11.6% OoV for in-domain corpus, drops to 1.3% with out-of-domain

- Our approach:
  - Consider infrequent a n-gram that appears less than t times in training
  - Select sentences from the out-of-domain corpora containing infrequent n-grams present in the test set

$\Rightarrow$ Informativeness of the selected sentences increases

On-line Sentence Selection for Infrequent n-grams Recovery

- Score each sentence s from the out-of-domain corpora using

$$\mathrm{f}(s) = \sum_{0<i<j<|s|} \begin{cases} \max\{0, t - \mathrm{N}(s_i^j)\} & \text{if } s_i^j \text{ appears in the test set} \\ 0 & \text{otherwise} \end{cases}$$

  where s_i^j is the n-gram of the sentence s from position i to j

- Pick the n sentences with the maximum score
- After selecting each sentence, update sentence scores
- Combine with probabilistic selection to avoid disturbing in-domain distribution

Bayesian adaptation for model stabilization

- Log-linear models typically estimated by means of MERT
- MERT turns unstable if amount of development data small

$\Rightarrow$ Apply Bayesian adaptation for stabilizing model weights:

  - In Bayesian adaptation, model parameters are viewed as random variables
  - Decision rule for training data T and adaptation data A:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}|\mathbf{f}; T, A)$$

with

$$p(\mathbf{e}|\mathbf{f}; T, A) = \mathcal{Z}' \int p(A|\Lambda; T) p(\Lambda|T) p(\mathbf{e}|\mathbf{f}, \Lambda) \; d\Lambda$$

Experiments: corpora provided

TED corpus (in sentences):

		$\|S\|$	$\|W\|$	$\|V\|$
train	En	47.5K	747.2K	24.6K
	Fr	47.5K	792.9K	31.7K
indev	En	571	9.2K	1.9K
	Fr	571	10.3K	2.2K
ofdev	En	641	12.6K	2.4K
	Fr	641	12.8K	2.7K

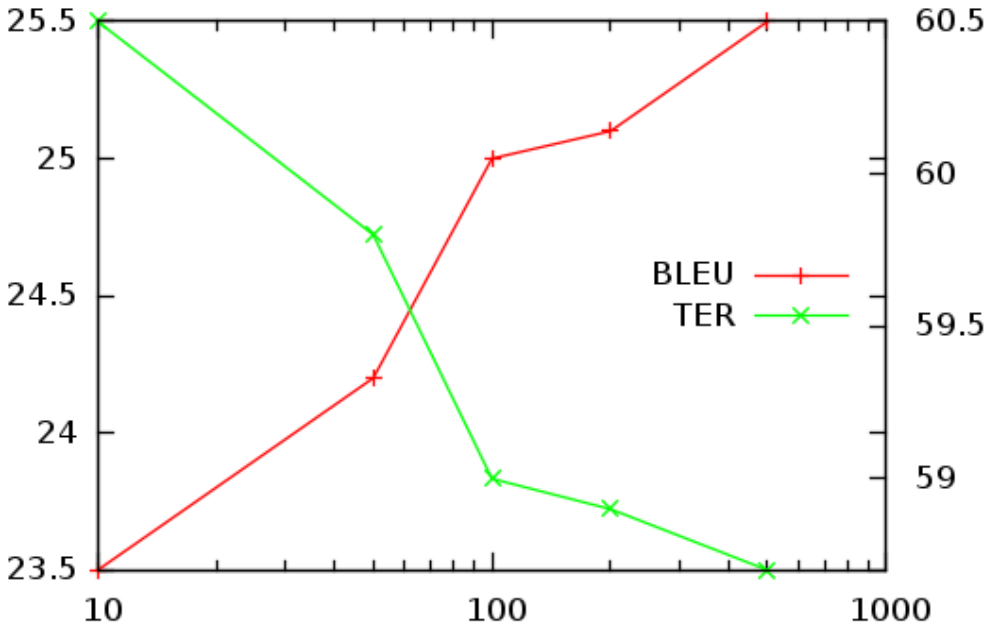
Out-of-domain corpora:

		$\|S\|$	$\|W\|$	$\|V\|$
Europarl	En	1.25M	25.6M	81.0K
	Fr	1.25M	28.2M	101.3K
News Commentary	En	67.6K	1.4M	35.6K
	Fr	67.6K	1.6M	43.3K
United Nations	En	5.0M	94.4M	302.7K
	Fr	5.0M	107.4M	283.7K
Gigaword	En	15.5M	302.9M	1.6M
	Fr	15.5M	360.6M	1.6M

Experiments: results

- Probabilistic Sentence Selection:

nK	BLEU	TER
0	23.2	60.8
10	23.5	60.5
50	24.2	59.8
100	25.0	59.0
200	25.1	58.9
500	25.5	58.7



Experiments: results

- Online sentence selection for infrequent n-grams recovery:

nK	t	$\|S\|$	MERT
50	-	96.9K	24.2/59.8
	1	99.9K	23.7/60.6
	10	101.9K	24.1/60.5
	25	101.9K	24.1/60.3
100	-	146.9K	25.0/59.0
	1	149.8K	24.6/59.6
	10	156.9K	24.1/60.2
	25	156.9K	24.6/59.4

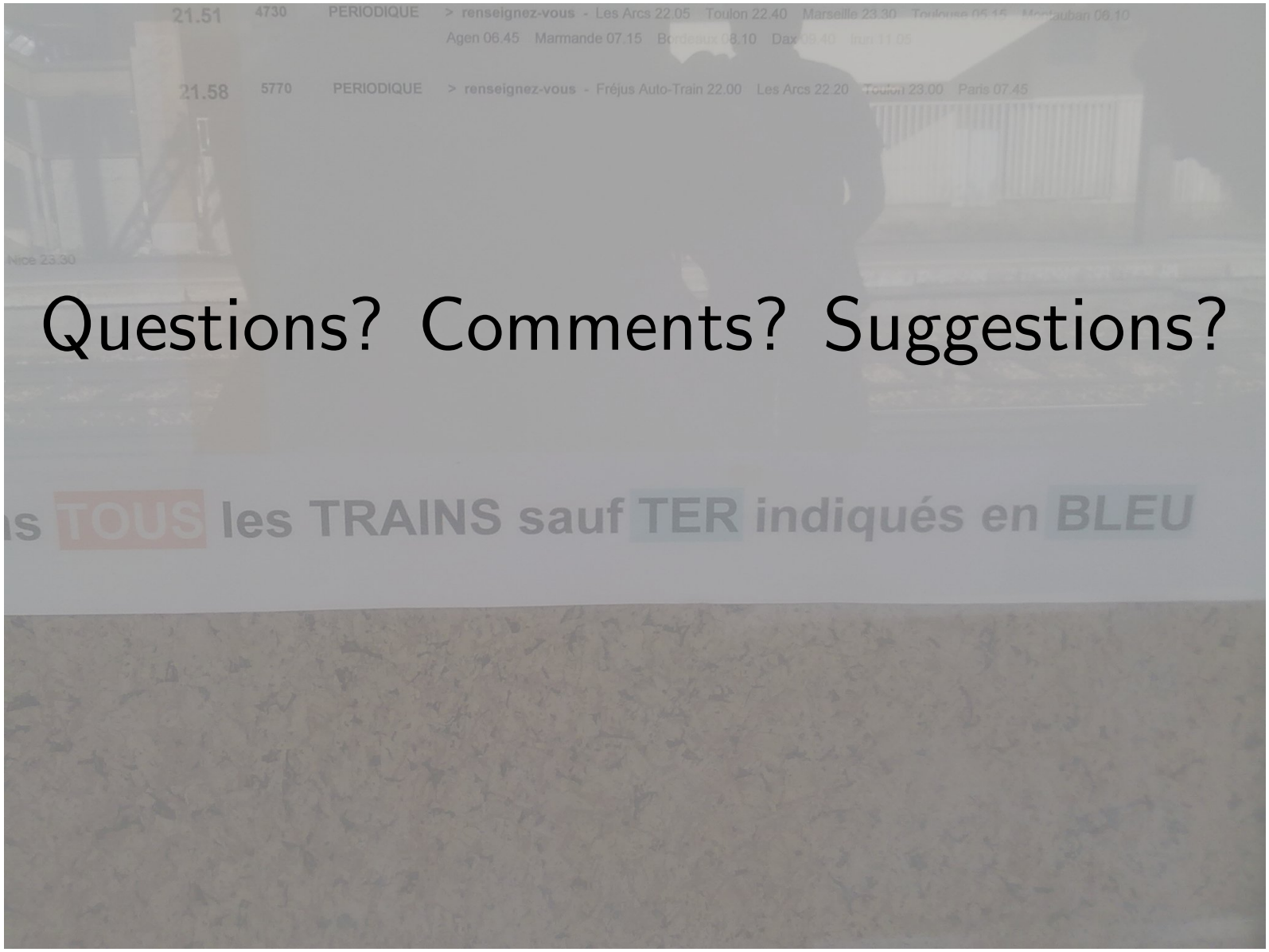
Experiments: results

- Online sentence selection for infrequent n-grams recovery:

nK	t	$\|S\|$	MERT	bayes
50	-	96.9K	24.2/59.8	24.7/58.7
	1	99.9K	23.7/60.6	24.9/58.8
	10	101.9K	24.1/60.5	25.2/58.4
	25	101.9K	24.1/60.3	25.2/58.4
100	-	146.9K	25.0/59.0	25.1/58.6
	1	149.8K	24.6/59.6	25.3/58.5
	10	156.9K	24.1/60.2	25.4/58.3
	25	156.9K	24.6/59.4	25.5/58.4

Conclusions and Future Work

- Conclusions:
  - Intelligent selection of training data seems to be a good strategy
  - Good results are obtained by using only a small percentage of the training sentences
  - Bayesian adaptaptation provides stability to results obtained
- Future work
  - Compare versus random data selection
  - Optimize log-linear combination in probabilistic sentence selection
  - Use sentence length normalization for the infrequent n-grams recovery technique
  - Adjust proportion of sentences used when combining both techniques


Questions? Comments? Suggestions?