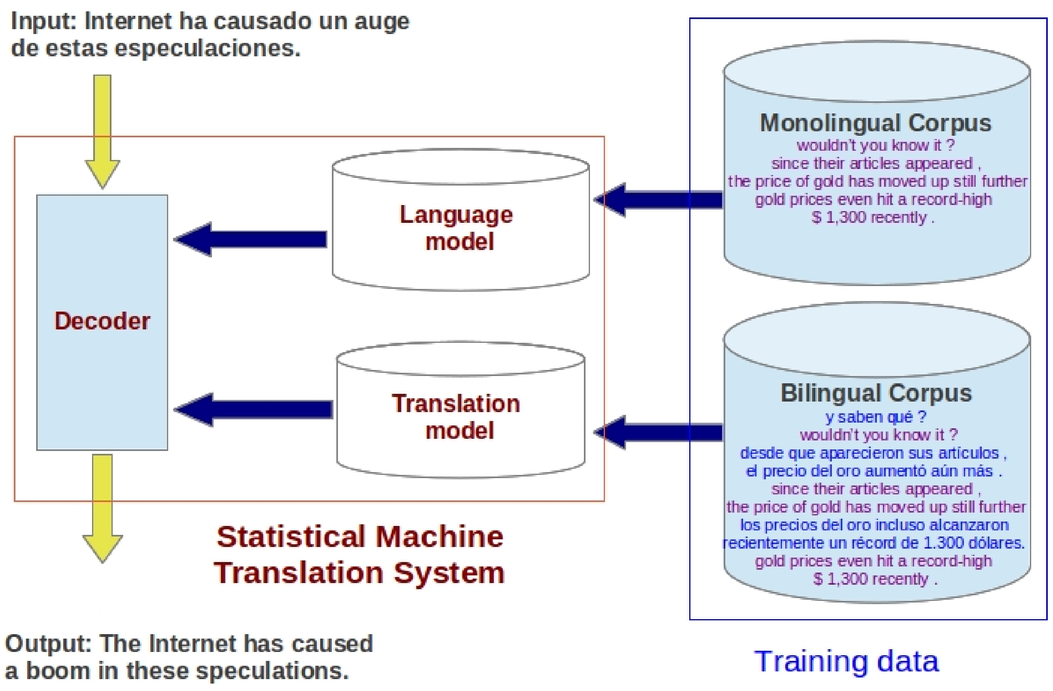
# Sentence selection in statistical machine translation

**Author: Mara Chinea-Rios***
**Directors: Francisco Casacuberta* and Germán Sanchis-Trilles***
PhD program Computer Science
*PRHLT Research Center {machirio,fcn,gsanchis}@prhlt.upv.es

## Introduction

- Statistical Machine Translation (SMT) system quality depends on the available training data
- Important factors: the size and the domain
- This work is focused in studying different strategies of *Bilingual sentence selection*

## Statistical Machine Translation




SMT principal equation

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y})$$

- $\mathbf{x}$ input sentence and $\mathbf{y}$ output sentence
- $Pr(\mathbf{y})$ Language model and $Pr(\mathbf{x}|\mathbf{y})$ Translation model

## Domain Adaptation Approaches

- Domain adaptation methods categories: corpus level and model level
- Corpus level approaches:
  - ↪ Select training data
  - ↪ Corpus weighting
  - ↪ Model combination
  - ↪ Latent semantics

## Infrequent n-grams recovery

- Main idea: increasing information of in-domain corpus
- *Infrequent n-grams*: An n-gram is considered infrequent when it appears less times than a given infrequency threshold $t$
- The infrequency score $i(\mathbf{x})$ is defined as:

$$i(\mathbf{x}) = \sum_{\mathbf{w} \in X} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w}))$$

  - $N(\mathbf{w})$ the counts of $\mathbf{w}$ of source language out-of-domain corpus
  - $C(\mathbf{w})$ the counts of $\mathbf{w}$ of source language in-domain corpus

## Cross-entropy selection

- Main idea: scoring sentences of out-of-domain corpus by cross-entropy
- The cross-entropy score of x is then defined as

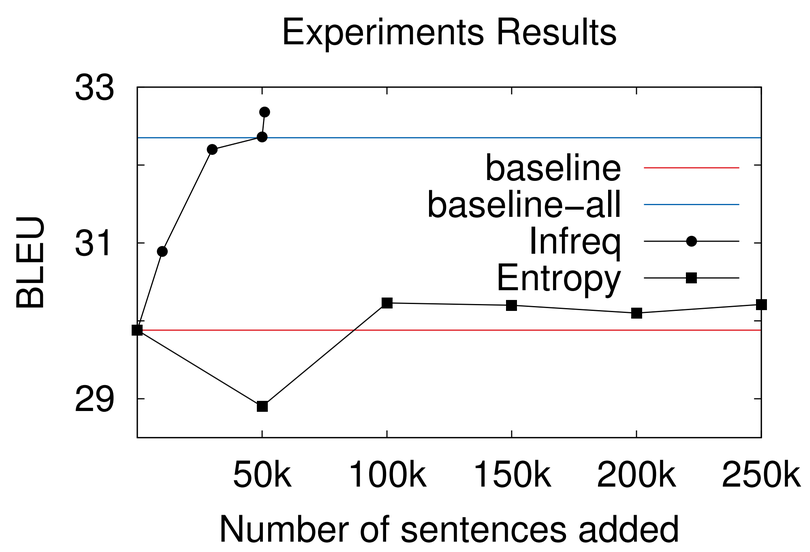
$$c(\mathbf{x}) = H_I(\mathbf{x}) - H_G(\mathbf{x})$$

  - $I$ be an in-domain corpus and G be an out-of-domain corpus
  - $H_I(\mathbf{x})$ be the cross-entropy, according to a LM trained on I
  - $H_G(\mathbf{x})$ be the cross-entropy, according to a LM trained on G

## Experiments

- Experimental setup:
  - Out-of-domain corpus: French-English of the Europarl corpus
  - In-domain corpus: EMEA corpus
  - Test: Medical test corpus 2014
  - Initial weights estimated with MERT on 2014 WMT dev. sets
  - Evaluation by means of BLEU
- Compared the selection methods with two baseline systems
  - **baseline-emea:** Training with EMEA corpus
  - **baseline-all:** Training with Europarl corpus ∪ EMEA corpus

## Best Results




| Strategy | BLEU | Number of sentences |
|---|---|---|
| Baseline-emea | 29.9 | 1.0M |
| Baseline-all | 32.4 | 1.0M + 1.4M |
| Cross entropy | 30.2 | 1.0M + 150k |
| Infreq | 32.7 | 1.0M + 51k |

- All strategies improve over baseline-nc from the very beginning
- Results very similar using less sentences of out-of-domain corpus.

## Conclusions

- Data selection has been receiving an increasing amount of interest within the SMT research community
- Data selection techniques obtain positive results using only a small fraction of the training data.

## Acknowledgements

Work supported by the European Union 7th Framework Program (FP7/2007-2013) under the Casmacat project (grant agreement nº 287576). Also funded by the Generalitat Valenciana under grant Prometeo/2009/014.