Learning to Translate: a statistical and computational analysis

Marco Turchi, Tijl De Bie, Nello Cristianini

University of Bristol (UK)

Department of Engineering Mathematics

Outline

- Motivation
- Introduction
- Experimental Setup
- Experiments
- Conclusion and Considerations

Motivation

- A belief in SMT is that “more data → better translation”.
- But:
  - how much parallel text do we need to obtain acceptable translation?
  - Do we have a constant increase in performance when adding more data?
  - If we have an exhaustive amount of parallel data, can the SMT model be a limitation?
  - Can we find the current limitation of the SMT approach?
- Some helpful facts:
  - data availability (Europarl, Hansard, UN corpus, Web, …);
  - recent advances in software (Moses, …);
  - computing power (HPC cluster, cloud computing, …).

Motivation

- **Extensive study** of a Phrase based SMT system using Moses, Europarl and a HPC cluster.
- Try to answer the previous questions by extrapolating the **performance of the system under different conditions**:
  - constantly increasing the training;
  - changing the system parameters;
  - adding noise to the system parameters;
  - …
- Investigate the **potentials and limitations** of the current technology analysing a STM system as a Learning System.
- **Explore** new aspects of a SMT system under a machine learning point of view.
- **Confirm** some previous results in SMT field.
- Suggest some possible **research directions**.

Introduction

- Performance of a learning system is result of (at least) two effects:
  - representation power of the hypothesis class:

    how well the system can approximate the target behaviour;

  - statistical effects:

    how well the system can estimate the best element of the hypothesis class.

Introduction

- They interact, with richest classes being better approximators of the target behaviour, but requiring more training data to identify the best hypothesis.

- In SMT, learning task is complicated by the fact that the probability of encountering new words or expressions never vanishes.

Introduction

- These observations lead us to analyze:
  - learning and unlearning curves;
  - flexibility of the representation class;
  - stability of the model;
- Experiments:

1. role of training set size on performance on new sentences;
2. role of training set size on performance on known sentences;
3. role of phrase length in translation table;
4. model perturbation: analysis and unlearning curves.

Experimental Setup

- Software
  - Moses.
  - Giza++: IBM model 1, 2, 3, and 4 with number of iterations for model 1 equal to 5, model 2 equal to 0, model 3 and 4 equal to 3.
  - SRILM: n-gram order equal to 3 and the Kneser-Ney smoothing algorithm.
  - Mert: 100 the number of nbest target sentence for each develop sentence.
  - Training, development and test set sentences are tokenized and lowercased.
  - Maximum number of tokens for each sentence in the training pair is 50.
  - TMs were limited to a phrase-length of 7 words. LMs were limited to 3.

Experimental Setup

- Data
  - Europarl Release v3 Spanish-English corpus.
  - Training set: 1,259,914 pairs.
  - Test and Development sets 2,000 pairs each.

- Evaluation Scores
  - BLEU, NIST, Meteor, TER, Unigram Recall, Unigram Precision, FMean, F1, Penalty and Fragmentation.
  - BLEU is used as evaluation score after we observed its high correlation to the other scores on the corpus.

Experimental Setup

- Hardware
  - University of Bristol cluster machine, http://www.acrc.bris.ac.uk/acrc/hpc.htm.
    - 96 nodes each with two dual-core opteron processors.
    - 8 Gb of RAM memory per node (2 Gb per core).
    - SilverStorm Infiniband high-speed connectivity throughout for parallel code message passing.
    - General Parallel File System (GPFS).
    - Storage 11 terabytes.
    - Torque v2.1.6p17 as the Resource Manager.
    - Maui v3.2.6p16 as the scheduler.

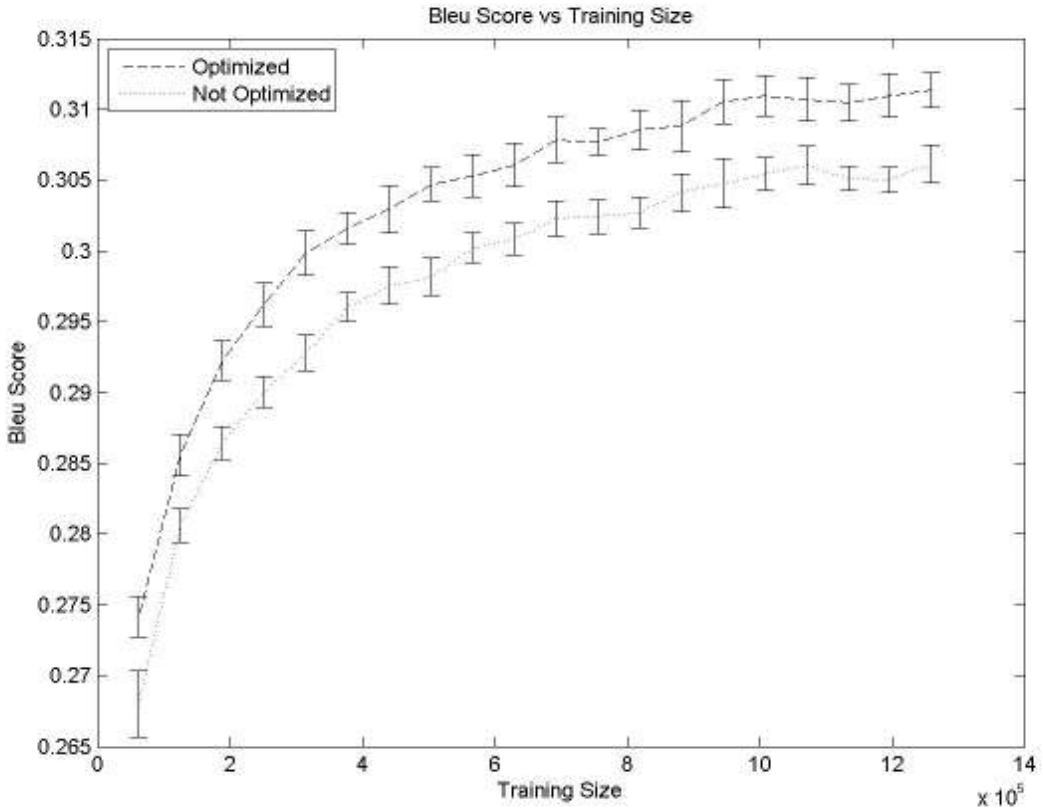
Role of training set size on performance on unknown sentences

- Analyse how performance is affected by training set size, by creating learning curves.
- Create subsets of the complete corpus by sub-sampling sentences from a uniform distribution, with and without replacement;
  - with replacement: analyse the performance on different training sets of the same size and the effects of optimization phase;
  - without replacement: study the SMT learning curves in the Linear-Linear and Linear-Log scales.

Role of training set size on performance on unknown sentences

- Create subsets of the complete corpus by sub-sampling sentences from a uniform distribution, **with replacement**.
- Ten random subsets for each of the 20 chosen sizes (each size 5%, 10%, etc of the complete corpus).
- For each subset, a new instance of Moses has been created.
- Development and test sets contain 2,000 pairs each.
- The experiments have been run for the models with and without the optimization step.

Role of training set size on performance on unknown sentences



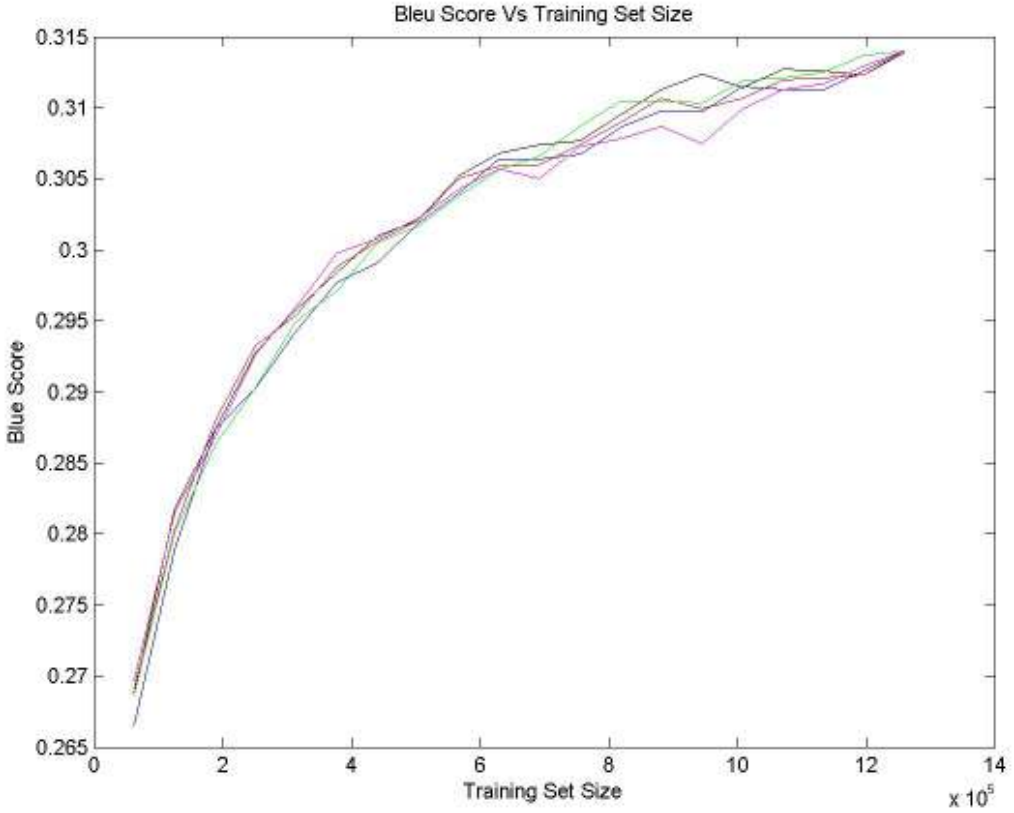
1. Small error bars.
2. Benefits optimization phase.
3. Curves affected by the Birthday paradox.

Role of training set size on performance on unknown sentences

- The whole training set is split in 20 blocks containing 5% of the data **without replacement**.
- Each increment of the training set size is a concatenation of a new block of data with the previous.
- Five random splits have been done of the whole training set.
- Each split produces a learning curve.
- A region of confidence is created between the learning curve with best performance and the learning curve with worst performance.

Role of training set size on performance on unknown sentences

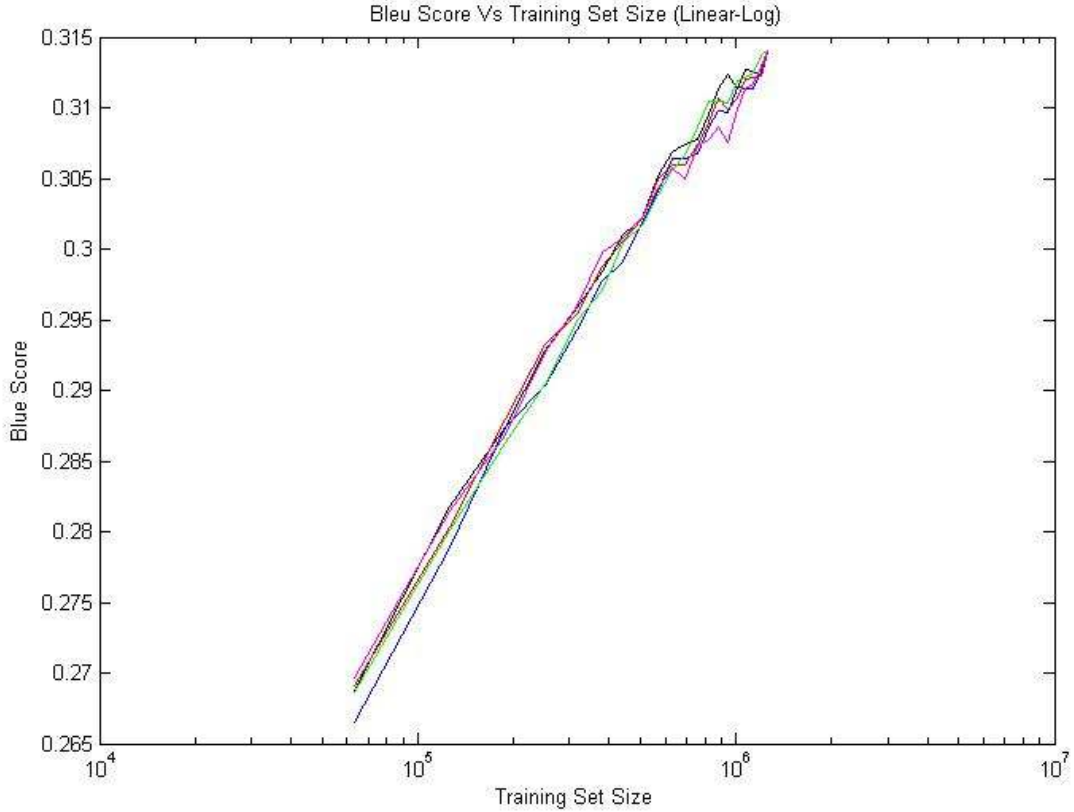
- Learning Curve region in Linear-Linear Scale.



1. Addition of massive amounts of data result into smaller improvements.

Role of training set size on performance on unknown sentences

- Learning Curve region in Linear-Log Scale.

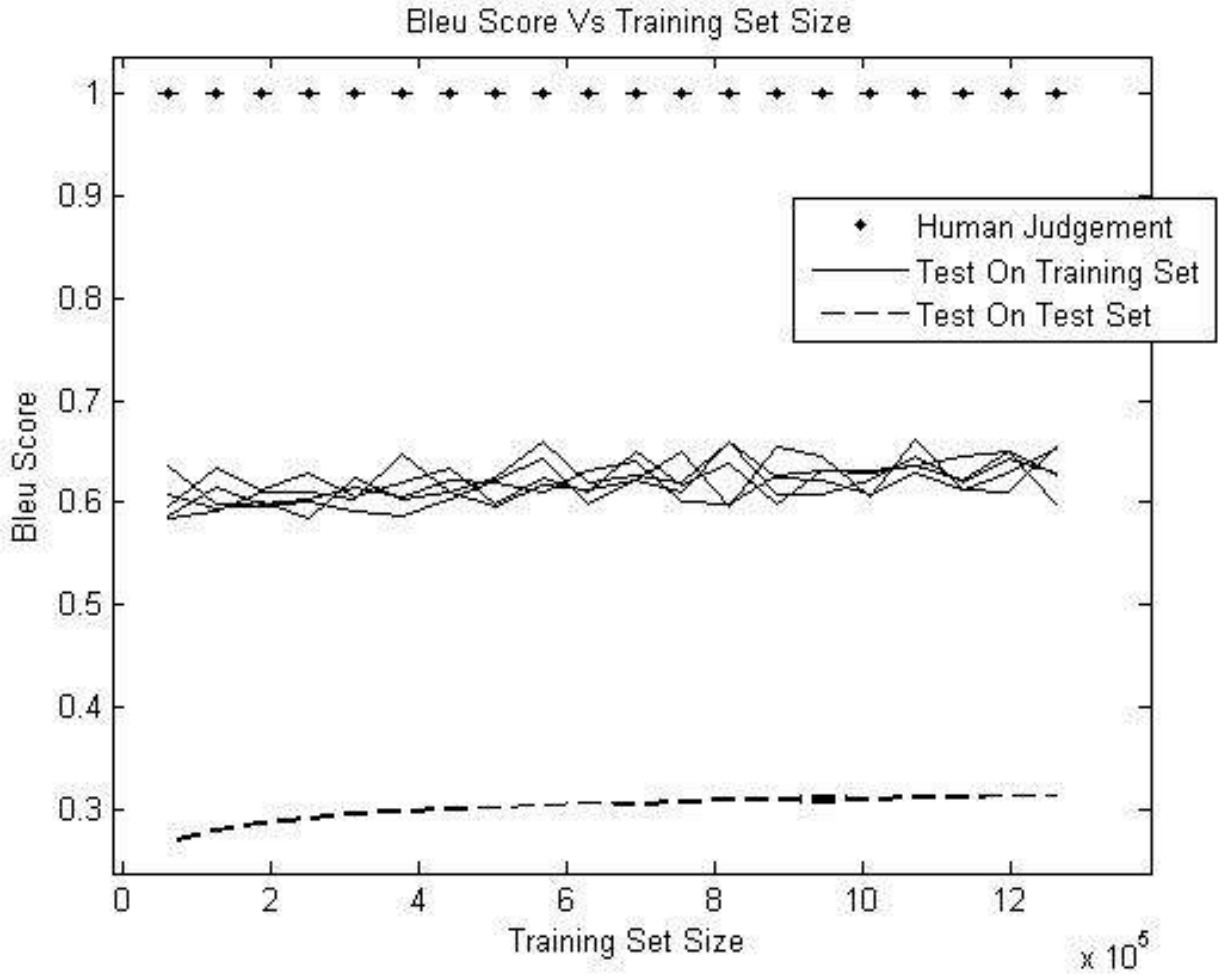


1. Logarithmic behaviour can not be excluded.
2. Learning curve is “logarithm at best”.

Role of training set size on performance on known sentences

- Experiment much like the one described above.
- Key difference: the test set was selected randomly from the training set (2,000 pairs after cleaning phase).
- An upper bound on the performance achievable by this architecture if access to ideal data was not an issue.
- Performance on translating training sentences are not due to simple memorization of the entire sentence.
- ”Human Translation” identifies the curve obtained using the reference sentences as target sentences.

Role of training set size on performance on known sentences



Role of training set size on performance on known sentences

- Fit a line to the test on test set learning curves in the linear-log scale using least squares.
- The approximated learning curve will reach the test on training learning curve with “only” 10^15 sentence pairs. It means:
  - 10^9 times the Europarl dataset
  - 3*10^9 years of proceedings of the European Parliament.
  - The Indexed Web contains at least 27.1 billion pages (Saturday, 09 May, 2009) by http://www.worldwidewebsize.com. If we assume that each page has 10 sentences, it would not be enough.

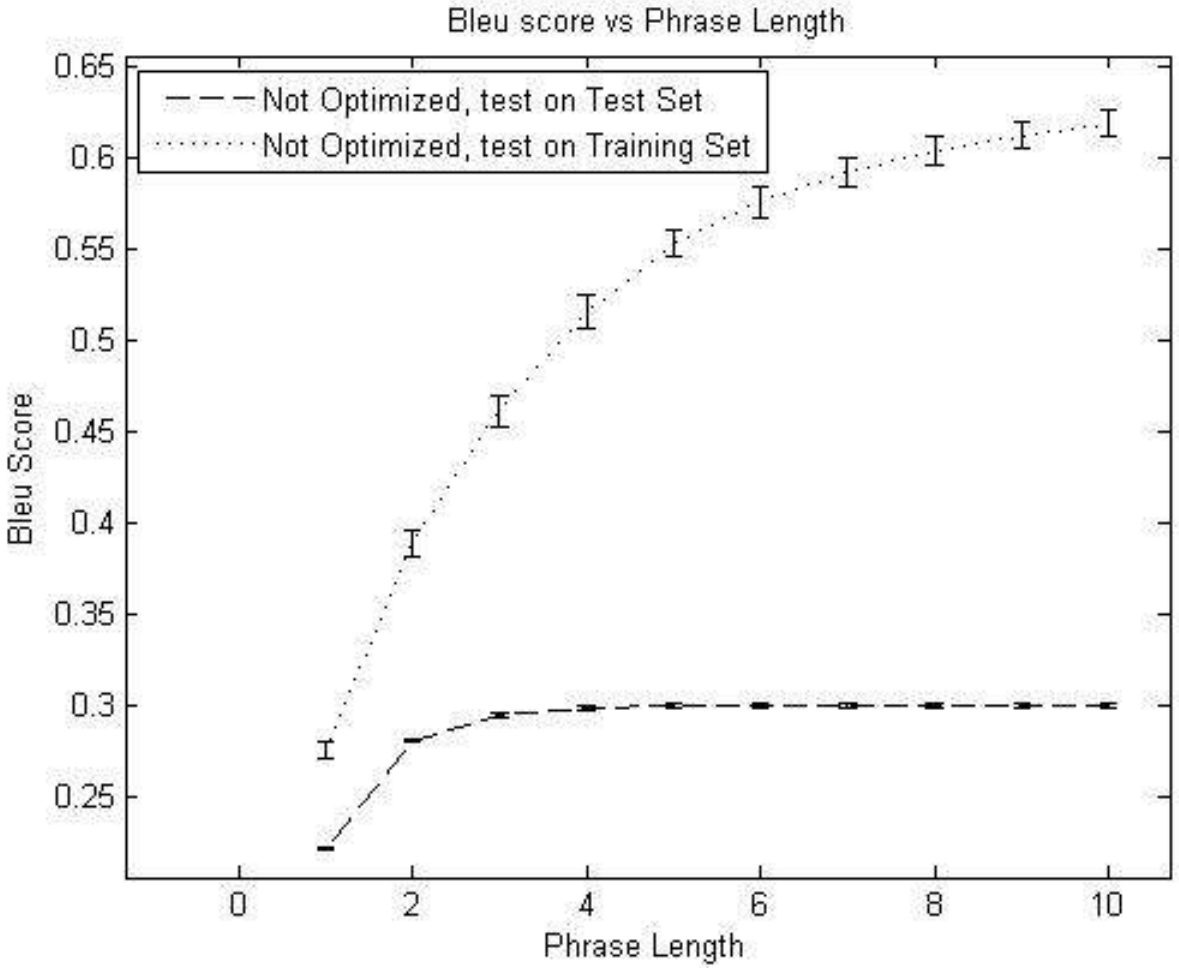
Role of training set size on performance on known sentences

- If the right information has been seen, the system can reconstruct the sentences rather accurately.
- System can represent internally a good model of translation.
- It seems unlikely that good performance will ever be inferred by increasing the size of training datasets in realistic amounts.
- Process with which we learn the necessary tables representing the knowledge of the system is responsible for the performance limitations.

Role of phrase length in translation table

- Gap between performances on training and on test sets is typically affected by model selection choices.
- Choice of the phrase length is crucial in the selection of the right model.
- Ten random subsets of the complete corpus containing 629,957 pairs of sentences have been created.
- For each subset, ten instances of the SMT have been created.
- Each instance has been trained using a different phrase length, from 1 to 10.
- Each model has been tested on the test set, 2,000 sentences, and on a random subset of 2,000 sentence from the training set.

Role of phrase length in translation table



Role of phrase length in translation table

- In both the learning curves there is a big improvement moving from the word by word translation, phrase length equal 1, to the phrase based model.
- No significant advantages seem to be present when phrase length is bigger than 4 in the “test on test set” learning curve.
- The rise of the phrase length improves the performance of the system when it has been tested on sentences sampled by the training set.
- Phrase length changes the dimension of the translation tables, but the system continues to prefer short phrase to long ones during the decoding phase

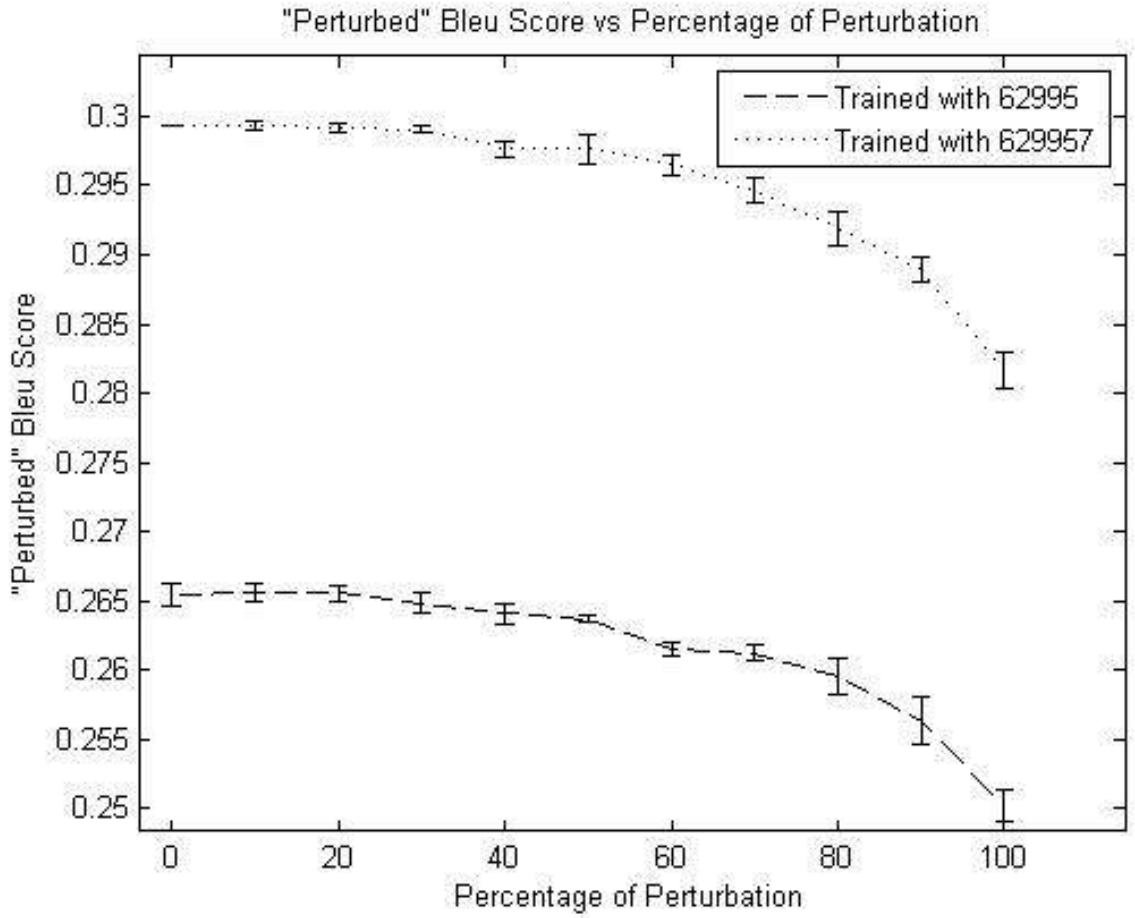
Model perturbation: analysis and unlearning curves

- The training step results in various forms of knowledge: translation table, language model and lambda parameters from the optimization.
- The internal models learnt by the system are lists of phrases, with probabilities associated to them.
- In order to simulate the effect of inaccurate estimation of the statistical parameters, two different experiments have been run:
  - a percentage of noise has been added to each probability in the LM and TM (*Adding Noise*);
  - noise has been added in the form of wrong associations between numerical and textual parts of LM and TM (*Randomization of Parameters*);

Model perturbation: analysis and unlearning curves

- Two models trained with 62,995 and 629,957 pairs of sentences have been used.
- Different value of percentage of noise between 0 and 1 have been used.
- The noisy probability is obtained as p' = min(1, p +v),where v = rand($-p \times k$, $+p \times k$) with percentage of noise $k \in [0, 1]$.
- If this quantity is bigger than one it has been approximated to one.
- For each model, for each value of *k*, ten experiments have been run.
- Optimization step has not been run.

Model perturbation: analysis and unlearning curves

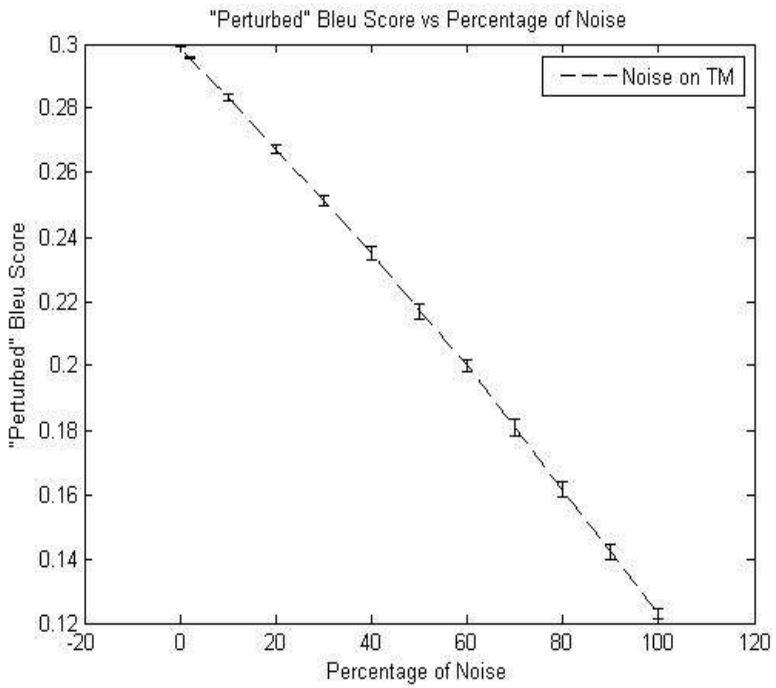


Model perturbation: analysis and unlearning curves (Randomization of Parameters)

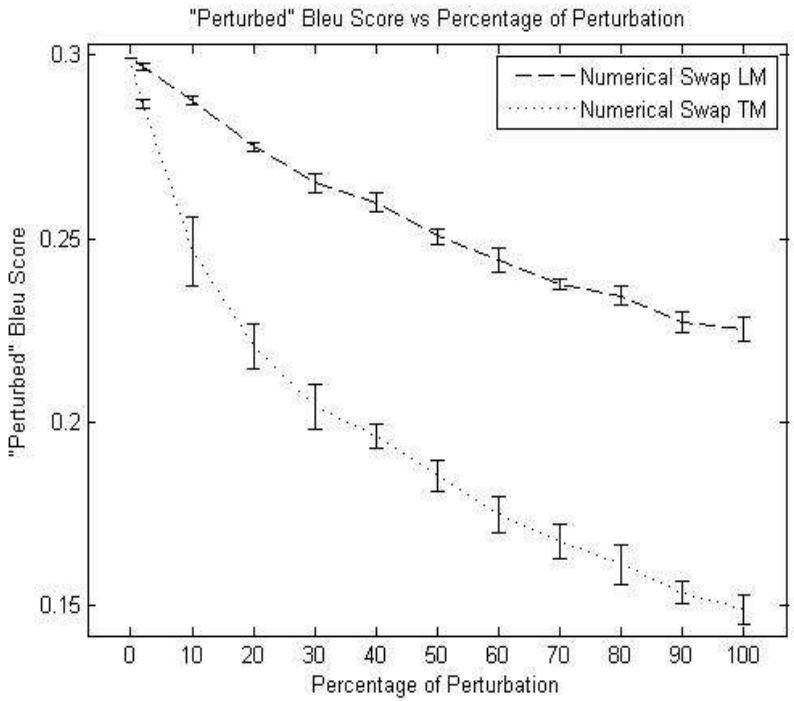
- We define:
  - Numerical Swap: given two entries of LM or TM, the numerical parts are swapped.
  - Words Swap: given two entries of TM, the target language phrases are swapped.
- Percentage of noise represents a certain number of swaps.
- Three different sets of experiments have been run:
  - Words Swap of TM;
  - Numerical Swap of LM;
  - Numerical Swap of TM.

Model perturbation: analysis and unlearning curves (Randomization of Parameters)

Words Swap in TM



Numerical Swap in TM and LM



Model perturbation: analysis and unlearning curves

- Adding Noise: gentle decline of the unlearning curve suggests that fine tuning of parameters does not seem to control the performance.

- Randomisation of Parameters: more aggressive noise produces more significant decline in performance. LM is less affected than TM. In Word Swap experiments a more rapid decline should be expected, but high redundancy in the TM prevents it.

Conclusion/Considerations

- Our experiments suggest that:
  - the current bottleneck is the **lack of sufficient data**, not the function class used for the representation of translation systems.
  - Adding more **i.i.d. data** does not appear to be a practical way to significantly improve performance.
  - The perturbation analysis suggests that improved **statistical principles are unlikely to make a big difference** either.
  - More than the accurate estimation of parameters, it is the **compilation of the translation tables** that drives the performance of the system.
  - Model selection choices, **phrase length**, will not make a big difference
  - Since it is **unlikely that sufficient data will be available by simply sampling a distribution**, one needs to address a few possible ways to transfer large amounts of knowledge into the system.

Conclusion/Considerations

- A research programme naturally follows from our analysis:
  - an effort to identify or produce datasets on demand.
    - It breaks the traditional i.i.d. assumptions on the origin of data.
    - It would also require an effective way to do confidence estimation on translations, to identify those instances where there is low confidence in the output.
  - Introduction of significant domain knowledge in the form of linguistic rules, to dramatically reduce the amount of data needed to essentially reconstruct them by using statistics.
- The barrier to improving performance is a direct consequence of Zipf's law and the frequency of phrases in text.
  - The impossibility of the algorithm to deal with unknown phrases, and their non-vanishing frequency in natural corpora conspire to create a fundamental limitation.

Learning to Translate: a statistical and computational analysis

Marco Turchi, Tijl De Bie, Nello Cristianini

University of Bristol (UK)

Department of Engineering Mathematics