An Intelligent Agent that Autonomously Learns how to Translate

Marco Turchi, Tijl De Bie, Nello Cristianini

Pattern Analysis and Intelligent Systems Group

University of Bristol (UK)

Outline

- Motivation
- The Agent
- Experimental Setup
- Stability Analysis
- Experiments
- Conclusion and discussion

Motivation

- A learning system deals with the following problems:
  - labelled training data are an expensive resources in terms of human interventions;
  - in complex tasks, it is not able to autonomously improve their performance capability providing new labelled training data.
- The idea is to design autonomous systems that can teach themselves (self-learning) how to perform a complex task.

Motivation

- Why Statistical Machine Translation (SMT)?
- Statistical Machine Translation has all the necessary aspects for one such agent to be non trivial:
  - a complex mapping between input and output (sentences in two different languages);
  - a reliable theory of learning (Phrase Based model);
  - clear criteria to assess performance (Bleu Score);
  - the availability of vast amounts of multilingual text that can be useful for training (newspapers produce huge amounts of articles every day).

Motivation

- Labelled data for SMT:
  - parallel sentences: one sentence is literally the translation of the other.

English	Italian
I declare resumed the session of the European parliament adjourned on Friday 17 December 1999.	Dichiaro ripresa la sessione del parlamento europeo, interrotta venerdi' 17 Dicembre 1999.
(the house rose and observed a minutes silence)	(il parlamento osserva un minuto di silenzio)

Motivation

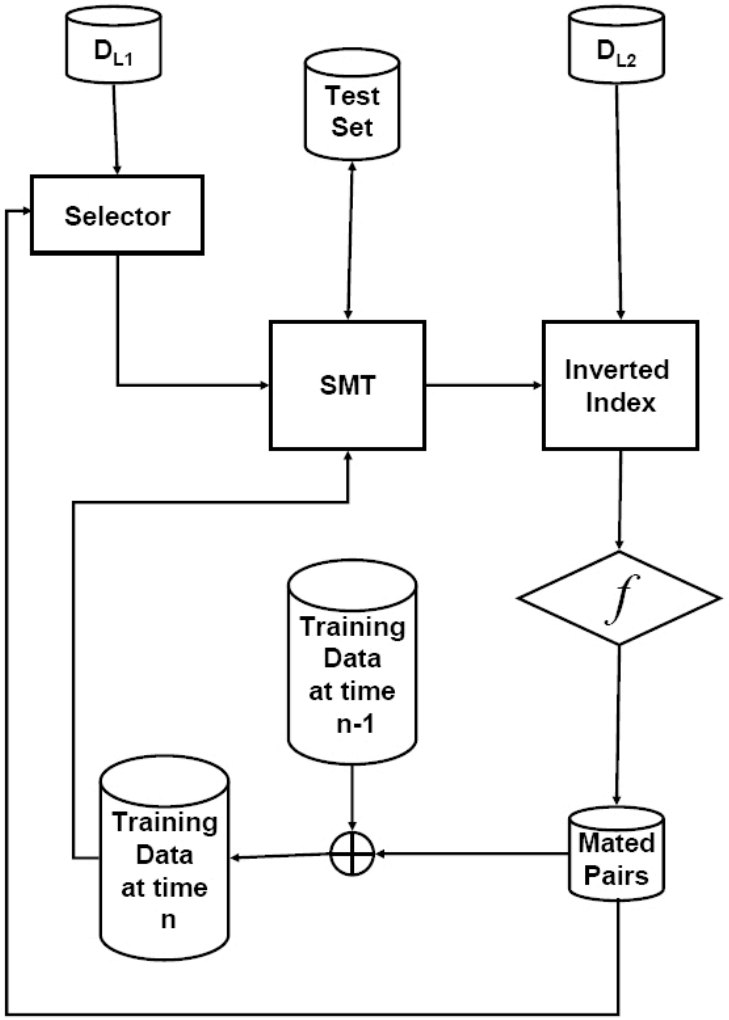
- The idea behind our agent is:
  - sentences coming from two monolingual datasets are mated by a pivot sentence that is the translation of the sentence in the source language (source sentence) to the target language (target sentence).
- Our agent is made of three components:
  1. a Statistical Machine Translation (SMT) system that translates the source sentences;
  2. a Cross-language information retrieval (CLIR) part that associates the translated sentence with a target sentence;
  3. a Tuning function that takes a decision on the quality of the association.

The Agent

Algorithm 1 Agent Algorithm

Require: $D_{L_1} \neq \emptyset$, $D_{L_2} \neq \emptyset$, a SMT trained with a parallel corpus T_0, α, f.

Create an Inverted Index (I) using D_{L_2}
while $D_{L_1} \neq \emptyset$ **do**
  $S_{L_1} \Leftarrow k$ randomly selected sentences from D_{L_1}
  for $s_{L_1} \in S_{L_1}$ **do**
    $mt \Leftarrow \text{SMT}(s_{L_1})$
    Find a set of possible similar sentences S_{L_2} of mt querying I
    if $f(S_{L_2},\alpha,s_{L_1}) ==$ True **then**
      Add (s_{L_1},s^*) to M, where s^* is the best element of S_{L_2} according to f
    end if
  end for
  $T_i \Leftarrow M$ and the training data T_{i-1} at previous iteration
  Train the SMT using T_i
  Test the performance of the new model
  Remove M from D_{L_1}
end while



Experimental Setup

- Data:
  - Europarl Release v3 Spanish-English. We split it to:
    - initial training set T_0: 125,441 pairs. Used for training the initial SMT model;
    - CLIR validation set: 12,544 pairs. Used to tune the agent parameters;
    - test set: 12,544 pairs. Used to test the algorithm;
    - mate set: 1,103,885 pairs. Used as a pool of unpaired sentences to combine to improve the SMT training set.
  - News corpus:
    - contains Spanish and English titles, and a short description of news of a multi lingual on-line newspaper for fourteen weeks;
    - for each week, more than 100 pairs are used as a test set. These pairs are obtained by mating news titles using information contained in the URL of each news item.

Tuning Function *f*

- A crucial role, because it takes a decision on the quality of the association.
- Takes as input:
  - a set of possible similar retrieved sentences S_{L_2} of m_t querying the index. Each sentence is associated with its similarity score to the query provided by the index;
  - a threshold α;
  - the source sentence s_{L_1}
- Output:
  - True: if the best retrieved sentence s* in S_{L_2} can be mated to the source sentence s_{L_1}.

Tuning Function *f*

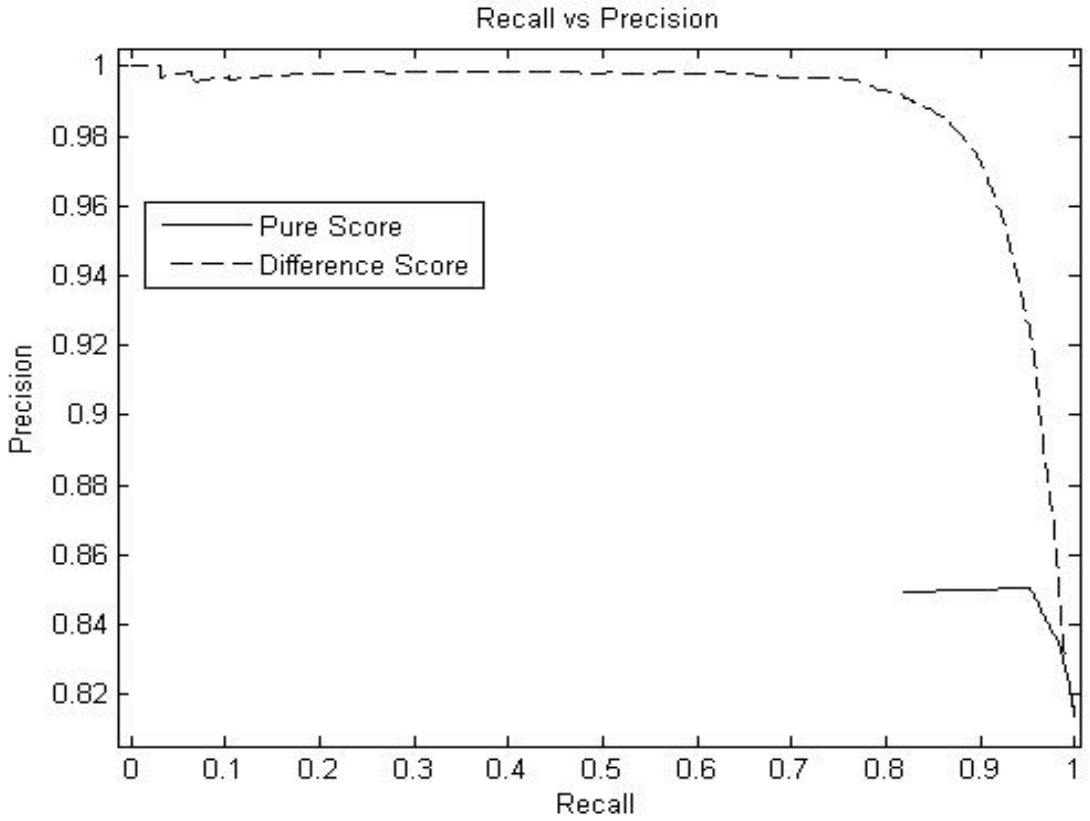
- The judgement of *f* is based on different contributions:
  - a metric that evaluates the similarity scores of the retrieved sentences;
  - a controller that analyzes whether the best retrieved sentence s* has not yet been mated;
  - a controller that checks the length of the best retrieved sentence s* and the source sentence (can not differ by more than 9 words).
- Note that *f* is tuned in a such way that it maximizes precision at the expense of recall (minimizes the number of incorrect mates).

Tuning Function *f*

- Analyse two different metrics:
  - the pure similarity score of the best retrieved sentence;
  - the marginal score: the difference between the similarity score of the best and the second best retrieved sentences. The idea is that s* can be mated to s_{L_1}, when the marginal score is big.
- Test the metrics:
  - build an index with the English part of the CLIR validation set;
  - translate the Spanish part of the CLIR validation set using a SMT system trained with T_0;
  - query each translated sentence to the index;
  - store the pure and marginal scores;
  - compute the ROC curve for both the metrics (for each Spanish sentence it is known which is the right mate in the index).

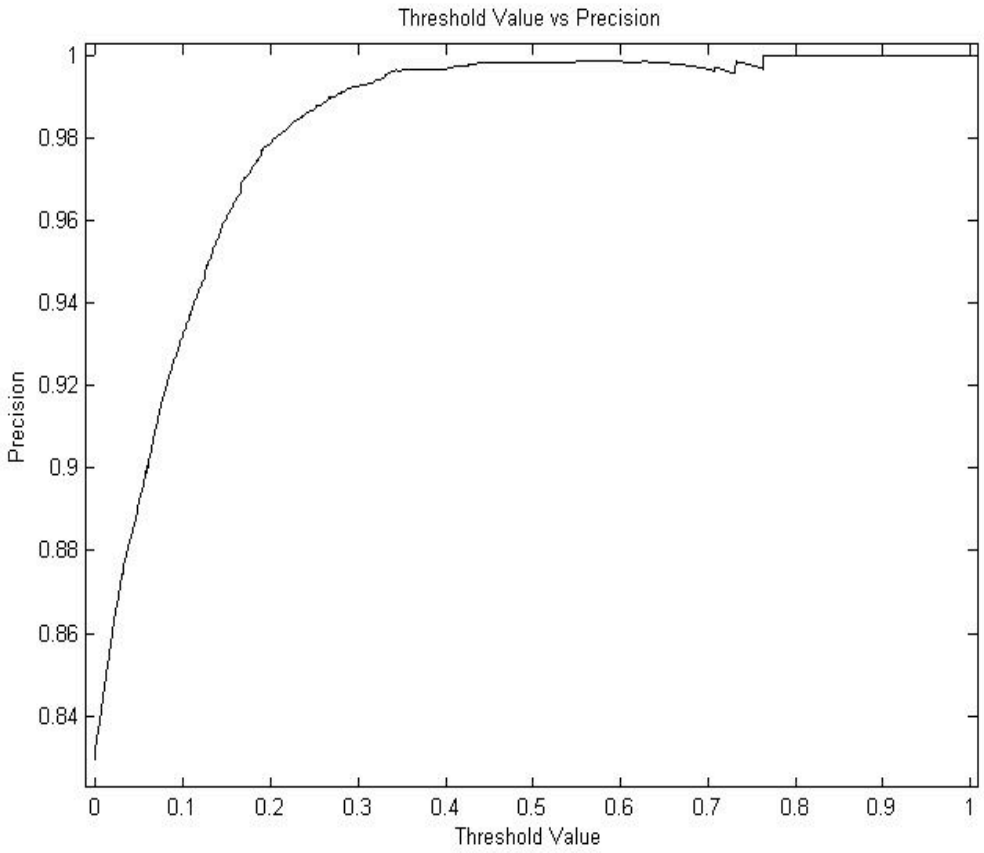
Tuning Function *f*

- Marginal score has better performance in terms of precision and recall.



Tuning Function *f*

- Chosen the metric:
  - select an acceptable precision value on the ROC curve;
  - select the relative threshold α from the Precision Vs Threshold value plot.
  - α = 0.5



Stability Analysis

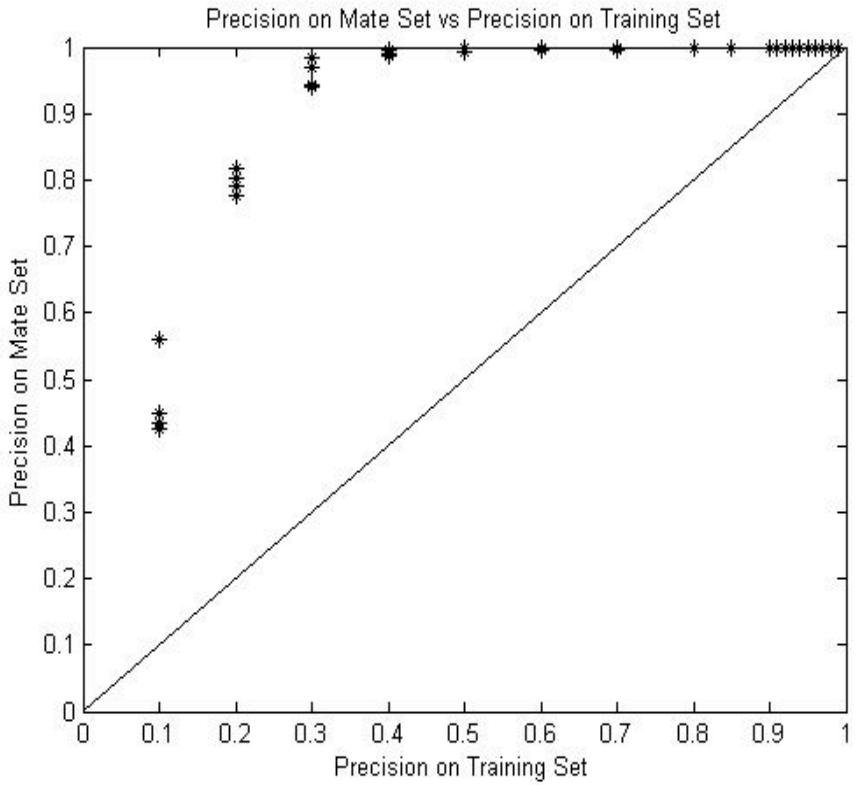
- All the conditions:
  - guarantee high precision;
  - do not avoid the presence of false positives.
- The addition of incorrectly mated pairs to the training set iteration after iteration may create a vicious circle.
- The relation between precision on the mate set (created set) and precision on the training set is what we call the **calibration curve**.
- Perform two analyses:
  - study of the calibration curves;
  - define a Stability Model.

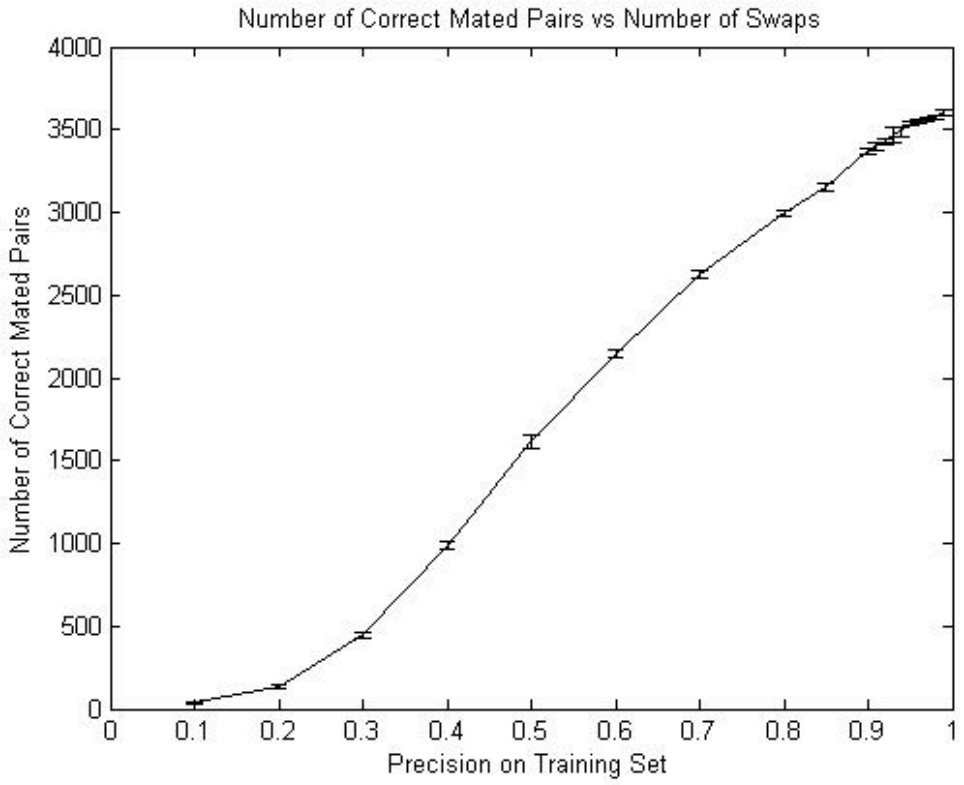
Stability Analysis: Calibration Curves

- To simulate low precision training data:
  - substitute a certain number of sentences in the English part of the training set (different percentages of noise, from 1% to 90% of the whole training set);
  - run the the agent using these training sets.
- For each experiment, we compute:
  - the number of correct mated sentences(*), precision on mate set(*) and precision on training set.

(*) These quantities are obtained comparing all the pairs mated by *f* with the ground-truth

Stability Analysis: Calibration Curves





Stability Analysis: Calibration Curves

- The agent is also able to extract correct mated pairs in the critical situation (with low precision on the training set):
  - with only 40% of correct training pairs the agent adds pairs that reflect roughly 99% precision on the “mate set”;
  - this is confirmed also by the number of correct mated pairs added, where the number of perfect mated pairs increases substantially after 40% precision on the training set.

- The calibration curves prove that the agent is:
  - stable and robust to noisy training data.
  - able to increase its knowledge starting from low precision training data.

Stability Analysis: Stability Model

- Define:
  - $K^{(t)}$: the number of correct mates in the training set at time t;
  - $N^{(t)}$: the number of sentence pairs in the training set at time t;
  - $k^{(t)}$: the number of correct mates added at time t;
  - $n^{(t)}$: the number of sentence pairs added at time t.
- Precision of the training set at time t (P), and precision added at time t (p) are respectively:

$$P^{(t)} = \frac{K^{(t)}}{N^{(t)}}; \qquad p^{(t)} = \frac{k^{(t)}}{n^{(t)}};$$

Stability Analysis: Stability Model

$$P^{(t+1)} = \frac{K^{(t)} + k^{(t)}}{N^{(t)} + n^{(t)}} = \frac{K^{(t+1)}}{N^{(t+1)}} = P^{(t)} - (P^{(t)} - p^{(t)}) \cdot \frac{n^{(t)}}{N^{(t)} + n^{(t)}}$$

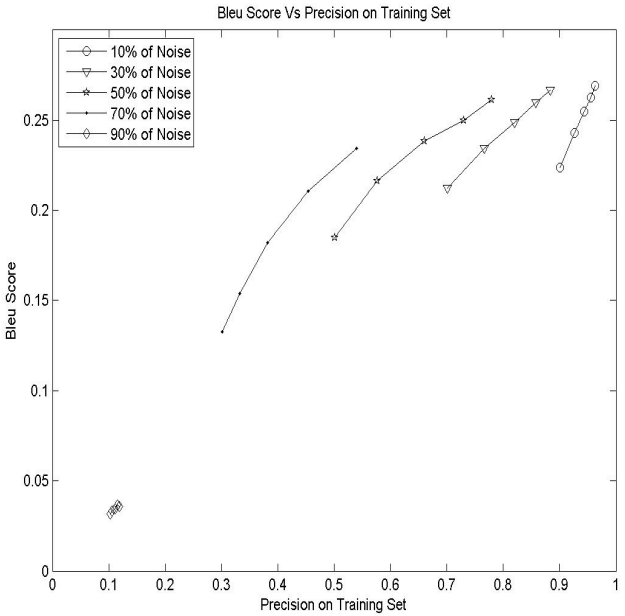
- $P^{(t)} - p^{(t)}$, an improvement factor, determines if there is an improvement (< 0) or decrement (> 0);
- $n^{(t)}/N^{(t)}+n^{(t)}$ a weighting factor, it is always between 0 and 1.
- $p^{(t)} = P^{(t)}$ is the bisection of the first quadrant in the previous precision on “mate set” vs precision on training set plot.
- If $P^{(t)} - p^{(t)} < 0$. This implies that $p^{(t)}$ has to be bigger than $P^{(t)}$.
- If sufficient information is in the starting seed our system can stably improve its performance.

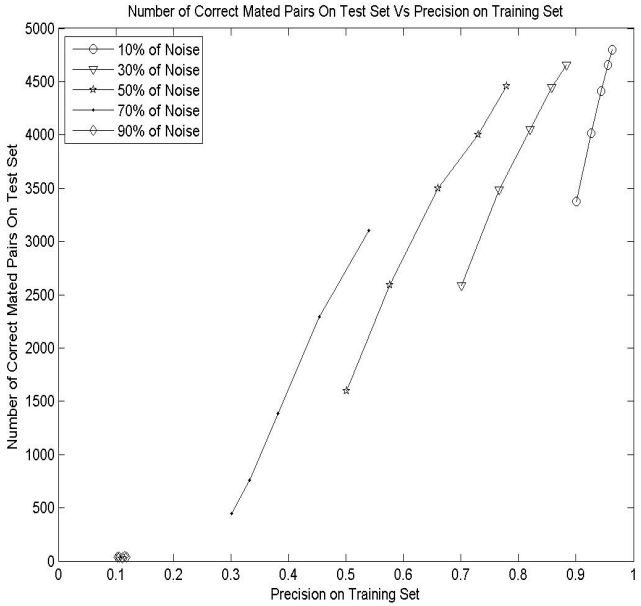
Experiments

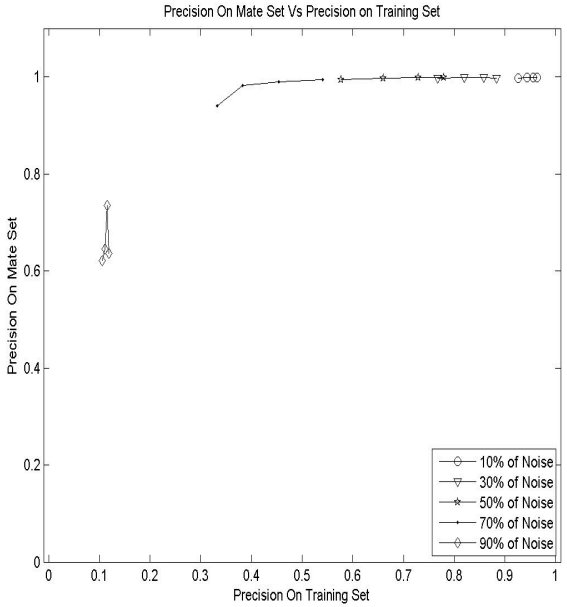
- Closed Environment:
  - Learning Curves with Low Precision Seeds: starting SMT system is trained with low precision seed.
  - Learning Curves with Pure Seeds: starting SMT system is trained with pure seed.
- Open Environment:
  - Web Analysis: test the agent on news data.

Learning Curves with Low Precision Seeds

- Analyze the behaviour of our agent letting it start from 5 different seeds with 10%, 30%, 50%, 70%, 90% of noise. Each seed contains 10,000 training pairs.
- Compute Blue score and number of correct mated pairs on test set and precision on “mate set”.





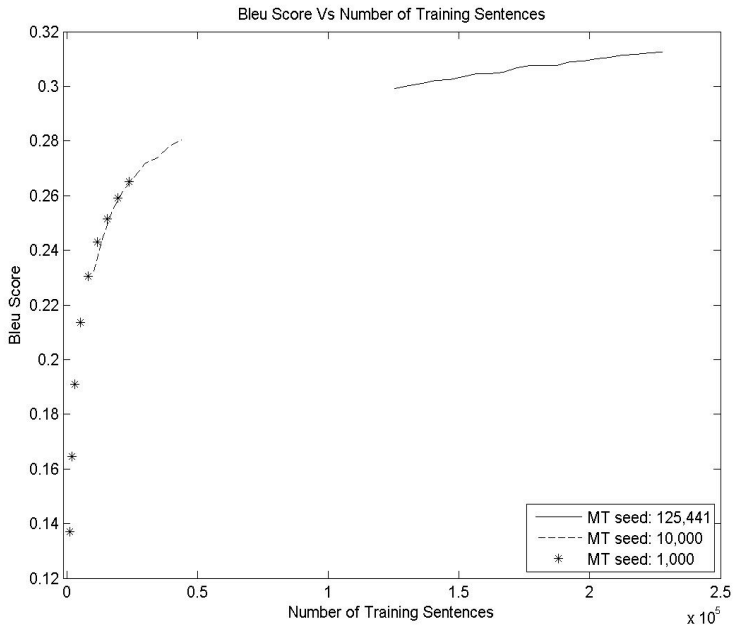


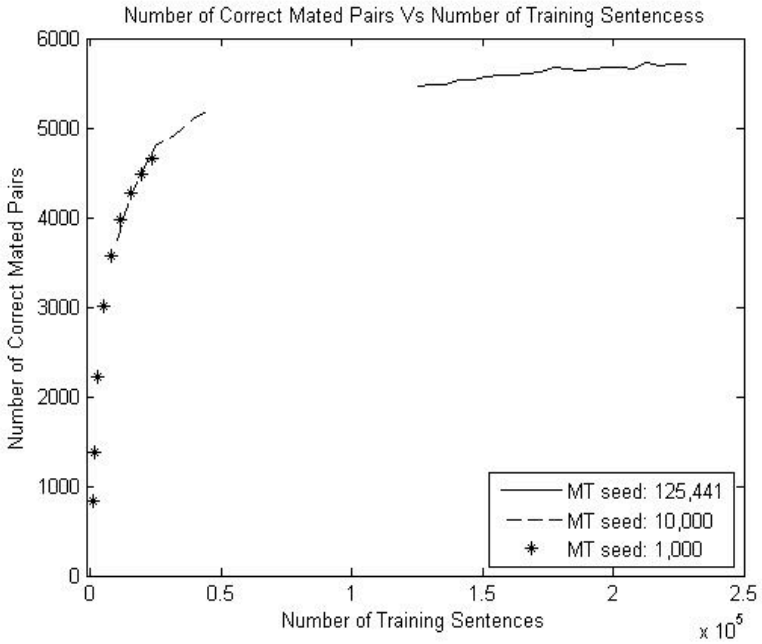
Learning Curves with Low Precision Seeds

- Starting with:
  - **very low precision seed**, the agent adds only few correct mated pairs. There is a slow increase in performance;
  - **higher precision seeds**, performance increases faster;
  - **70% noisy seed** the agent starts with small performance, but in few iterations reaches high precision on mate sets.
- The agent:
  - constantly increases its translation capability;
  - respects the stability condition of our model at each iteration;
  - works proficiently in noisy environments. It is robust to the number of incorrect mated pairs which can be added by mistake.

Learning Curves with Pure Seeds

- Analyze the behaviour of our agent letting it start from 3 different pure seeds:
  - T_0; 10,000 random selected pairs from T_0; 1,000 random selected pairs from T_0.
- All the metrics are computed on the test set.





Learning Curves with Pure Seeds

- Small amounts of training data are enough to start to increase the agent knowledge (1,000 or 10,000 are very small quantities).
- Increasing the seed dimension, improvement brought by each iteration decreases.
  - If the agent has poor knowledge, all the information that it obtains produces big improvements.
  - The more expert it becomes, the less any additional information introduces significant new knowledge.
- The agent constantly also continues to get advantage from new data when the number of sentences in the training set is sufficiently big.

Web Analysis

- Test the learning capability of the agent on Web data.
- The biggest problem of the web data is that the test set contains comparable sentences.

  Source: Huelga de profesores en Portugal

  Comparable_reference: Portuguese teachers protest over evaluation plans

  Correct_reference: Teachers' strike in Portugal

  Target: strike teachers in Portugal

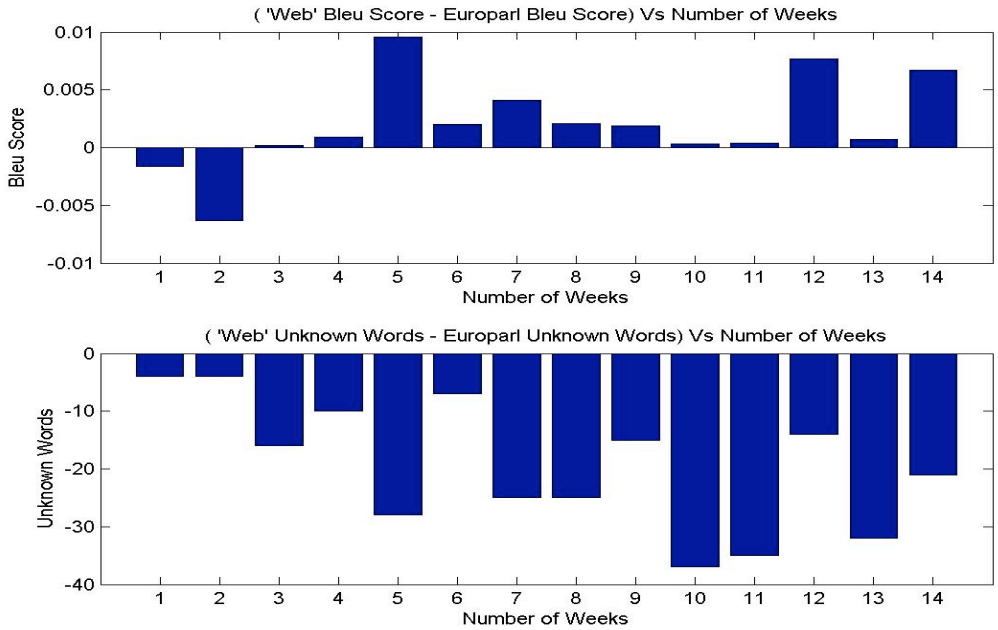
  - Blue score between a target sentence and a reference-compareble sentence is small, but target sentence is not necessary a bad translation.
- To better analyze the agent performance:
  - compare Bleu scores obtained using our agent with Bleu score obtained using random sampled pairs from the Europarl mate set.

Web Analysis

- A seed of 10,000 sentences is randomly sampled from T_0 and cloned twice to $T_{0_{web}}$ and $T_{0_{europarl}}$.
- Two different SMT systems are trained on $T_{0_{web}}$ and $T_{0_{europarl}}$
- We run our agent on the Web data, and at each iteration *i*:
  - a certain number, *x*, of parallel sentences from the Web news are mated by the agent and added to $T_{t-1_{web}}$;
  - a different model is created adding only *x* randomly selected from the Europarl "mate set" pairs to$T_{t-1_{europarl}}$;
  - The "Europarl" and "Web" models are tested on the test set of the week after computing Bleu score and unknown words.

Web Analysis

- For each week, performance using the Web and Europal models are obtained.
- Plot the difference between “Web” and “Europarl” Bleu score and number of unknown words.



Web Analysis

- Adding pairs mated by our agent produces more benefit than adding Europarl data.
- Bleu score is in general higher and the number of unknown words is smaller.
- Starting with a small and general purpose seed of training data, the agent is able to:
  - increase its translation capability;
  - learn new words or phrases;
  - in the presence of a different domain, adapt its knowledge to this new domain.

Conclusion and discussion

- We have investigated:
  - the **stability** of the agent;
  - the agent behaviour in low precision training set conditions, finding **high robustness** and stability;
  - a **theoretical model** has been developed to allow the agent to have proper cognition of its status at each time.
- The agent autonomously increases its **capability to learn new information** about languages.
- Although the addressed problem is within the context of SMT, this class of models will be **needed in many areas** of learning agents on the Web.

Conclusion and discussion

- In SMT, a crucial problem is not only the drift of languages, but also the drift of topics.
  - If a system is trained on medical training data, it is hard for it to translate sport documents. This is very common in news content.

- Our intention is to test the agent on the full Web using Web search engines to gather the two monolingual datasets.

Thanks for your attention!

An Intelligent Agent that Autonomously Learns how to Translate

Marco Turchi
marco.turchi@gmail.com
http://www.marcoturchi.com