AFRL-RI-RS-TR-2016-172

NON-PARAMETRIC MODEL DRIFT DETECTION

USC INFORMATION SCIENCES INSTITUTE

JULY 2016

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2016-172 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
ALEKSEY PANASYUK
Work Unit Manager

/ S /
MICHAEL J. WESSING
Deputy Chief, Information Intelligence
Systems and Analysis Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
JULY 2016	FINAL TECHNICAL REPORT	APR 2015 – APR 2016

4. TITLE AND SUBTITLE

NON-PARAMETRIC MODEL DRIFT DETECTION

5a. CONTRACT NUMBER FA8750-15-C-0071

5b. GRANT NUMBER N/A

5c. PROGRAM ELEMENT NUMBER N/A

6. AUTHOR(S)

Aram Galstyan

5d. PROJECT NUMBER IARP

5e. TASK NUMBER MD

5f. WORK UNIT NUMBER DS

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2016-172

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The IARPA seedling effort explored an automated framework for model maintenance. The effort calculated in an unsupervised fashion the difference between the dataset that was used to train the model and the new dataset on which the model is to be applied (this is done using a new tool called CorEx that automatically estimates structure in high dimensional data through correlation) . The experimentation took place on datasets made up of text documents. The difference between datasets used to estimate potential error (drop in accuracy) that the model would incur if applied on the new dataset. The tradeoff between time cost of retraining the model and potential error of applying the original model on the new dataset will used in making the decision on whether to retrain or not.

15. SUBJECT TERMS

Ontology Extraction, Regulatory Compliance Assistant, Extraction of executable rules from regulatory text

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**ALEKSEY PANASYUK**
U	U	U	UU	38	**19b. TELEPHONE NUMBER (*Include area code*)** **N/A**

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

List of Figures

Summary

In this project, we developed and validated a novel methods for detecting and correcting model drift in unsupervised settings. The proposed approach has two components: drift detection, and drift correction. For the first sub-problem, we have utilized our recently developed method, Correlation Explanation, or CorEx, for detecting distributional changes in high dimensional data. For the second sub-problem, we have developed a decision-theoretic approach that provides a computational framework for trading off cost versus expected performance gain. We have validated the above framework on two tasks in NLP domain, topic modeling, and machine translation. Our main findings are summarized as follows:

- We can measure important distributional changes with CorEx using the notion of *surprise*. We also find that a decrease in classification accuracy is accompanied by increase in surprise, although the opposite is not always true: there are some distributional changes that result in increasing surprise, but not necessarily affecting the algorithmic performance.
- While an alternative measure of model drift (empirical KL distance) can sometime produce similar results, its behavior is less reproducible across the datasets. Also, there are scenarios where this measure will fail detect important distributional changes.
- The proposed drift-correction framework performed as expected, with some small variations across the datasets. We found that the optimal frequency of retraining depends on the cost of retraining, e.g., the higher the cost, the less frequent retraining. The main advantage of the proposed approach is its ability to adapt to different cost/benefit ratio for a given scenario.

Below we report on our main findings in more details.

Introduction

Most machine learning methods operate under the assumption that the training and the test data are sampled from the same distribution. Unfortunately, in most cases, this assumption does not hold. For instance, in the case of machine translation, a model learned using a large corpus of parallel-annotated data in one source domain (e.g., newswire) is employed to translate documents in a different domain (e.g., scientific literature) because of the difficulty in retraining the model for the target domain in a timely or cost-efficient manner. Furthermore, in most real-world situations the data generation process is itself time varying (e.g., even the news domain shifts over time and new words/phrases enter the vocabulary). Thus, it is important to have efficient and accurate methods for detecting, quantifying, and mitigating the negative consequences of model drift.

The goal of this effort was to develop and validate a computational framework for model drift detection and correction in unsupervised settings. In particular, the project was addressing the following two broad questions:

1. Given a reference dataset, and a model trained on that dataset, to what extent can we apply the learned model directly to a new dataset without retraining?
2. When a drift is detected, what is the optimal strategy of retraining the model, depending on the cost of retraining, expected performance deterioration if not retrained, and so on.

For the first sub-problem, we have utilized our recently developed method, Correlation Explanation, or CorEx, for detecting distributional changes in high dimensional data. For the second sub-problem, we have developed a decision-theoretic approach that provides a computational framework for trading off cost versus expected performance gain.

To validate our approach, we have focused on topic modeling and monitoring problem, with a particular emphasis on understanding and characterizing model drift in scientific literature. Our experiments were geared toward demonstrating the two central aspects of our approach: In the first set of experiments, we evaluated the ability of the proposed approach to detect and quantify model drift. And in the second set of experiments, we have performed a quantitative evaluation of the proposed decision-theoretic framework for drift correction, based on cost-sensitive model retraining paradigm. In addition to topic modeling, we have also conducted experiments in another domain, machine translation.

Methods, Assumptions, and Procedures

The proposed approach consists of two main components, *Measuring Drift* and *Decision Framework*, as schematically illustrated by the colored boxes in Fig.1. We now describe each individual component in more detail.

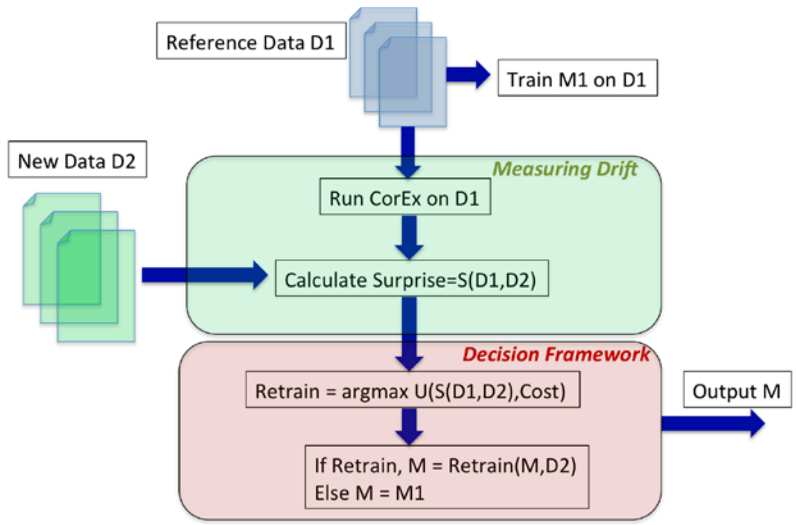


Figure 1 Schematic illustration of the proposed Model Drift detection & Correction Framework

Measuring Model Drift via Surprise

Consider a setting where we are given two datasets, and would like to know whether the model learned for the first dataset can be applied to the second dataset. In the absence of labeled data, one alternative for measuring model drift is to characterize the distance between distributions from which those datasets originate. For instance, one could compare the various moments of those distributions (e.g., skewness or kurtosis). A more general approach pursued here is to characterize the change in the distribution themselves, using information theory. Intuitively, distributional differences can be described using the metaphor/language of "surprise." The surprise of an observation, x, is defined as its negative log likelihood, $S(x) = -\log p(x)$ (according to the "true" distribution, $p(x)$).

Imagine we are given one or several samples from a new, unknown distribution, $q(x)$. Are these samples different enough from the original distribution that we should re-train our model? Here we suggest a model-free approach for calculating the surprise. Estimating information-theoretic quantities from samples is difficult because they depend on the unknown probability, $p(x)$. If x is actually an n-dimensional variable, then the number of samples needed to estimate $p(x)$ is exponential in n. Instead of estimating $p(x)$, we define an information-theoretic optimization whose output produces a function $f(x)$ that is an upper bound for the true surprise. Greater computational effort in the optimization leads to successively tighter bounds eventually converging to the true bound. This approach relies on the recently introduced method of Correlation Explanation (CorEx) that defines an information-theoretic coarse-graining for high-dimensional data [1,2]. CorEx is a fully non-parametric method that grounded in information theory, works as follows: Given a set of high-dimensional sample points, it learns a hierarchical generative model that explains the observed correlations in the covariates. Specifically, given the observed covariates, CorEx introduces a layer of hidden variables, so that, when conditioned on those variables, the covariates become uncorrelated (or less correlated). Mathematically, this is done by minimizing an information-theoretic entity called *Total* (conditional) *Correlation*; see [1,2] for more details.

Drift Correction Methods

Once we have detected a distributional shift, the next step is to decide whether to retrain the model or not. Our proposed drift correction framework is based on a utility-maximization approach. Namely, our decision process is formulated via the following optimization problem:

$$R = \underset{r=1,0}{\operatorname{argmax}} U(r)$$

$$U(r) = -Cr - \gamma Err(r)$$

Here C denotes the cost of retraining; γ is a parameter controlling the relative tradeoff between cost and error, and r is a binary variable indicating whether there is retraining or not: when $r = 1$, we retrain the model, otherwise we do not; and finally, $Err(r)$ is the expected error for the particular choice of r. Since we do not have a way of estimating the error (in the absence of labeled data), we will use empirically measured relationship between surprise and error. As detailed in previous reports, this relationship can be approximated by piecewise linear function.

In our experiments reported below, we used $\gamma = 1$, and will tried 5 different values for the cost C, to ensure that we capture various realistic scenarios.

For comparison, below we have considered the following baselines:

- B1: No retraining
- B2: Always retraining;
- B3: Retraining when the change in surprise is more than 10%.

In our experiments, we have compared those approaches across two different performance metrics: *utility*, as defined above, and *classification accuracy*; and *utility* as defined above.

Results and Discussions

We now describe the datasets used in our validation studies, and the main findings from our experiments.

Datasets

Topic Modeling Task

The experiments were conducted on three datasets, arxiv, PubMed, and NIPS.
The arxiv data contains paper abstract from different disciplines and sub-disciplines, including Computer Science, Math, Physics, covering the period 1995-2013. Here we will focus on CS papers, which itself is comprised of different subcategories, CS.AI, CS. Logic, etc. The PubMed dataset contains papers from four journals, *BMC Bioinformatics*, *BMC Developmental Biology*, *BMC Genomics*, and *BMC Cancer*. These papers span from 2001 to 2015. Finally, the NIPS dataset contains papers from NIPS (Advances in Neural Information Processing Systems) conference series from 1988-2003.

For all datasets, we set up a binary classification task, by dividing the papers into two classes, A and B. For the *arxiv* data, we considered papers in CS.AI as class A, and the rest of the CS papers as class B. For PubMed data, we considered *BMC Cancer* to be class A, and all the other papers as class B. For NIPS, we set up class A to contain all the papers on neural network and neuroscience, while the other papers constitute the class B. Note that we had to manually label NIPS papers for setting up this classification task. Additionally, for NIPS we also planned a different classification task, where class A contained papers written by a selected group of authors, and class B included all the other papers. Unfortunately, as indicated below, the classifier did not achieve a reasonable accuracy even for the reference dataset, so those experiments turned out to be not that valuable.
The statistics of the datasets are listed in the tables below.

NIPS data

Number of documents	2709
Dictionary size	4005
Number of authors	2484

PubMed data

Number of documents	19369
Dictionary size	23222
Number of journals	4

arxiv data

Number of documents	184015
Dictionary size	9989

Machine Translation Task

One of the main required resources for current state of the art MT systems is parallel data. The main idea behind our experiments is thus as follows: We assume we have a parallel data in one domain, but not in the second domain. Thus, when we train an MT engine in one domain, we should decide whether to apply it to a second domain, or to get additional parallel data from that domain and retrain. Since building MT engines is a time and resource consuming exercise, we have designed a careful plan for experimentation.

- **Data**: French-English parallel data from http://opus.lingfil.uu.se/
  - D1: OpenSubtitles2015 (66k/51M/338.5M docs/sentences/words)
  - D2: MultiUN (87k/13.2M/320M docs/sentences/words)
- **MT engines development**
  - Select training data: 20M words of training data per domain
  - 2,500 sentences for tuning per domain
  - Train 3 MT engines: D1, D2, D1+D2
- **Test data setup**
  - Select 5,000 documents for each domain (D1, D2)
  - Construct a test dataset D_{test} by taking a weighted combination of D1 and D2 (for different weights of each component).
  - Translate each document in D_{test} with each of the three engines.

The quality of the MT engine is measured by the Bleu score.

Drift Detection for Topic Modeling Task

Experiments with gradual shift

First, we look at the experiments with gradual drift. In this settings, we use papers published in $\{Y_1, Y_2,..Y_t\}$ for training, and then use each of the years $\{Y_{t+1}, Y_{t+2},..Y_T\}$ as training sets.

Her we focus on PubMed and NIPS datasets.

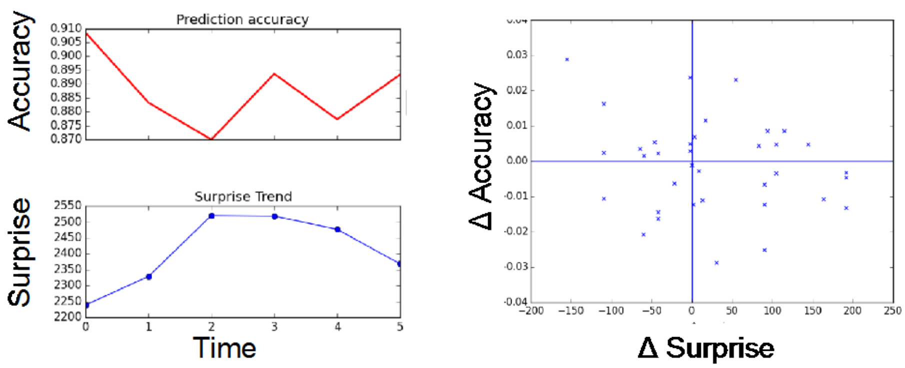


Figure 2 Temporal drift results for PubMed dataset

Fig. 2 shows results from a representative run for the PubMed data. We used all the papers published in the range 2001-2009 as the training set. Correspondingly, the papers published during 2010-2015 are the test set. The number of topics for this experiment is set to 50. After learning an LDA model on the training, or reference, set D_R, we use an SVM classifier that separates the classes A and B. We then apply this classifier to each publication year in the test set D_T, and track the prediction accuracy. We also calculate the surprise $S(D_R, D_T)$ for each of the testing dataset D_T.

In the left panel, we plot the prediction accuracy and surprise against time. We observe that the dip in accuracy is match by an increase in surprise. After the decrease, the accuracy fluctuates, while the surprise becomes almost constant, and then even decreases. On the right, we show a scatter plot of the change in accuracy vs change in surprise. Note that we have performed multiple runs for generating the scatter plot.
Next, we discuss results from he NIPS data, shown in Fig. 3, which shows a typical run with a number of topics set to 100. The papers from the first 8 conferences comprise the training set, and each subsequent conference is treated as a test set.

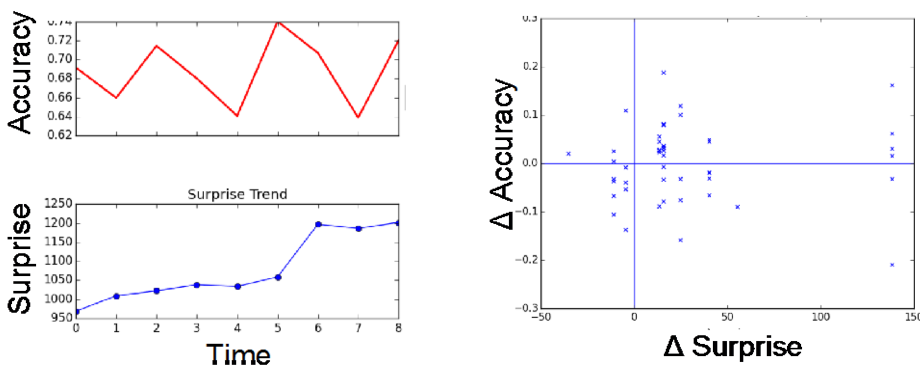


Figure 3 Temporal drift results for the NIPS dataset

We note that the classification accuracy does not show a clear temporal tendency to decline. Instead, it rather fluctuates around the value $Acc \approx 0.68$. The surprise, on the other hand, increases, except for the 5th and 8th test sets. This is somewhat counterintuitive, although we note that most of the increase in surprise is very moderate, except for the 7th test set, which also accompanies relatively big drop in accuracy. Also, the scatter plot on the right does not show any significant correlation between change in accuracy and change in surprise.

Finally, we consider the second classification task with NIPS dataset, where the goal is to classify the papers according to their authors. Namely, class A contains all the papers written by a selected list of K authors, whereas class B contain all the other papers. As we already mentioned, the results for this classification task were poor even for the reference dataset, as shown in Fig. 3. Thus, this particular problem is not very useful from the perspective of detecting model drift.

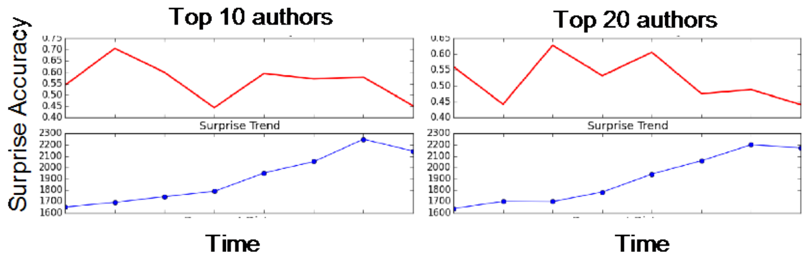


Figure 4 Author classification results for NIPS dataset

Experiments with abrupt shift

Now we focus on experiments when the model drift is abrupt. The abrupt shift was implemented as follows.

Let $D_A = \{a_1, a_2, .. a_N\}$ and $D_B = \{b_1, b_2, .., b_N\}$. be two corpora of documents for our binary classification task. For instance, in the case of NIPS data, D_A is the set of papers in the category NN (Neural Networks), whereas D_B is the set of papers in the other category NotNN (not Neural Networks). Furthermore, let $D_C = \{c_1, c_2, .. c_M\}$ be yet another set of papers. For instance, this can be a subset of the NotNN category papers. Or, it can be from a totally different collection.

We divide the sets D_A and D_B **randomly** intro a Reference and Test sets, $D_A = D_A(Ref) + D_A(Test), D_B = D_B(Ref) + D_B(Test)$. So now we have a Reference and Test datasets, $\boldsymbol{D_{Ref}} = D_A(Ref) \cup D_B(Ref)$ and $\boldsymbol{D_{Test}} = D_A(Test) \cup D_B(Test)$. The LDA model, the corresponding SVM classifier, and CorEx, will be trained on this set D_{Ref}. Note that according to the above construction, $\boldsymbol{D_{Ref}}$ and $\boldsymbol{D_{Test}}$ come from the same distribution. Thus, an SVM classifier trained on $\boldsymbol{D_{Ref}}$ should produce accurate results for $\boldsymbol{D_{Test}}$ as well.

We now introduce a parameterized abrupt drift as follows:

1. Let α be a number between 0 and 1.
2. For each document $\boldsymbol{d}$ in $\boldsymbol{D_{Test}}$ do the following:
    a. Select a random document **c** from set D_C
    b. For each word in document **d**, with probability α, replace it with a random word from document **c**
3. Repeat the above for $\alpha = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

For each value of α, the above procedure will result in a new, ***drifted*** test set $\boldsymbol{D_{Test}}(\alpha)$. For each of those dataset, we will test for model drift and calculate the relationship between accuracy and surprise.

In addition to surprise a calculated via CorEx, we will also consider another measure of distributional distance for measuring the drift. The KL distance between the Reference and Test datasets, $\boldsymbol{D_{Ref}}$ and $\boldsymbol{D_{Test}}$ is defines as follows,

$$KL(\boldsymbol{D_{Ref}} || \boldsymbol{D_{Test}}) = \sum_d p_{Ref}(d) log \frac{p_{Ref}(d)}{p_{Test}(d)}$$

where the summation is over all the possible documents (in bag of words representation), and p_{Ref}, p_{Test} are the distributions generating the reference and test sets, respectively.

Direct evaluation of KL distance is impossible due to the enormous state space. Thus, we replace the distributions p_{Ref}, p_{Test} by their empirical approximations as follows. We first combine all the documents in the Reference (Test) set into a single document, and corresponding bag of work representation, e.g., $\boldsymbol{BOW_{Ref}} = \{w_1, w_2, .., w_K\}$, where $K$is th dictionary size, and w_k is the number of times the k-th word appears in the corpus. Let $N = \sum_{k=1}^{K} w_k$ be the total number of words in the corpus, and let $x_k = \frac{w_k}{N}$. We then approximate p_{Ref} by multinomial distribution $Mult(x_1, x_2, .., x_K)$. The approximation for the test set is defined similarly. With this approximation, the KL distance cam be calculated easily.

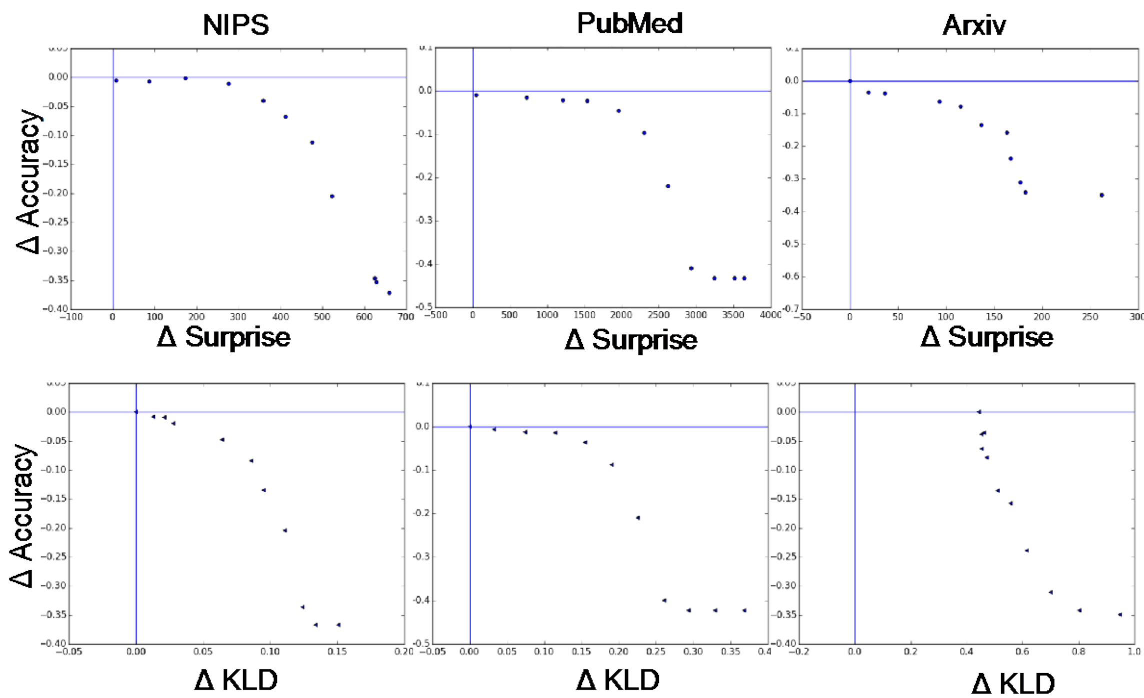


Figure 5 Relationship between change in accuracy and surprise/empirical KL distance

The results from the experiments are shown in Fig. 5, where we show a scatter plot of the change in accuracy $\Delta Accuracy$ vs change in surprise $\Delta Surprise$ (upper panel) and the empirical KL distance ΔKLD (lower panel). Each point corresponds to a specific value of α.

First, we observe that for the abrupt drift scenario, the relationship between the change in accuracy and surprise is less noisy, and more well-defined. Namely, if the change in surprise is larger than some threshold value, then there is also a noticeable drop in accuracy. The threshold value varies from dataset to dataset, which is expected. More importantly, the relationships are qualitatively similar for three datasets (despite quantitative differences).

We observe a similar picture with the empirical KL distance, especially for the NIPS and PubMed dataset. However, for the arxiv dataset (which has shorter documents), the behavior is more abrupt, which suggests that the empirical KL distance is not a universally good measure of distributional change.

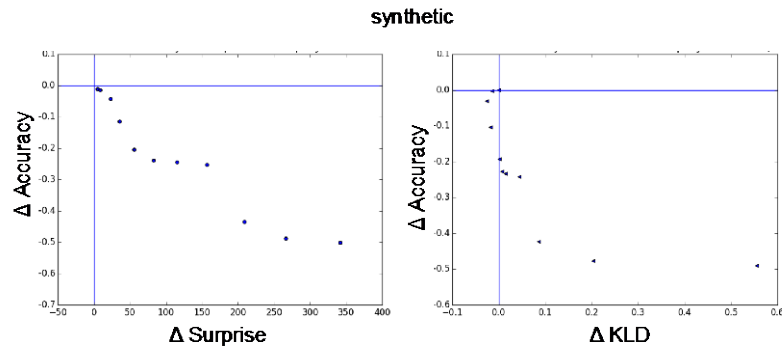


Figure 6 Relationship between change in accuracy and surprise/empirical KL distance for synthetic data

Indeed, our experiments with synthetic data confirm this point. For instance, Fig 5 shows results from experiments with synthetically generated data, which shows that the empirical KL distance is not detecting any change, even though the accuracy has dropped significantly. In fact, it is possible to construct example where the empirical KL distance fails to recognize distributional changes. For instance, let x_k^A and x_k^B be the probabilities of seeing the k-th word in class A and B, respectively. Since the empirical KL distance depends only on the aggregate probability $x_k^A + x_k^B$, any transformation of those probabilities that does not change the aggregate probability will not change p_{Ref} (or p_{Test}) either. The surprise, on the other hand, is calculated by first estimating the correlation structure of the data, and will detect any relevant distributional drift.

Drift Correction for Topic Modeling Task

For drift correction, we used NIPS, PubMed, and arxiv datasets for our experiments, and focused on abrupt drift scenario as described above. Recall that in this scenario, we have a ***drifted*** test set $\boldsymbol{D_{Test}}(\alpha)$ for each value of the mixing parameter . We will conduct our drift correction experiments for each of those datasets.

We start our discussion of results with the NIPS data.

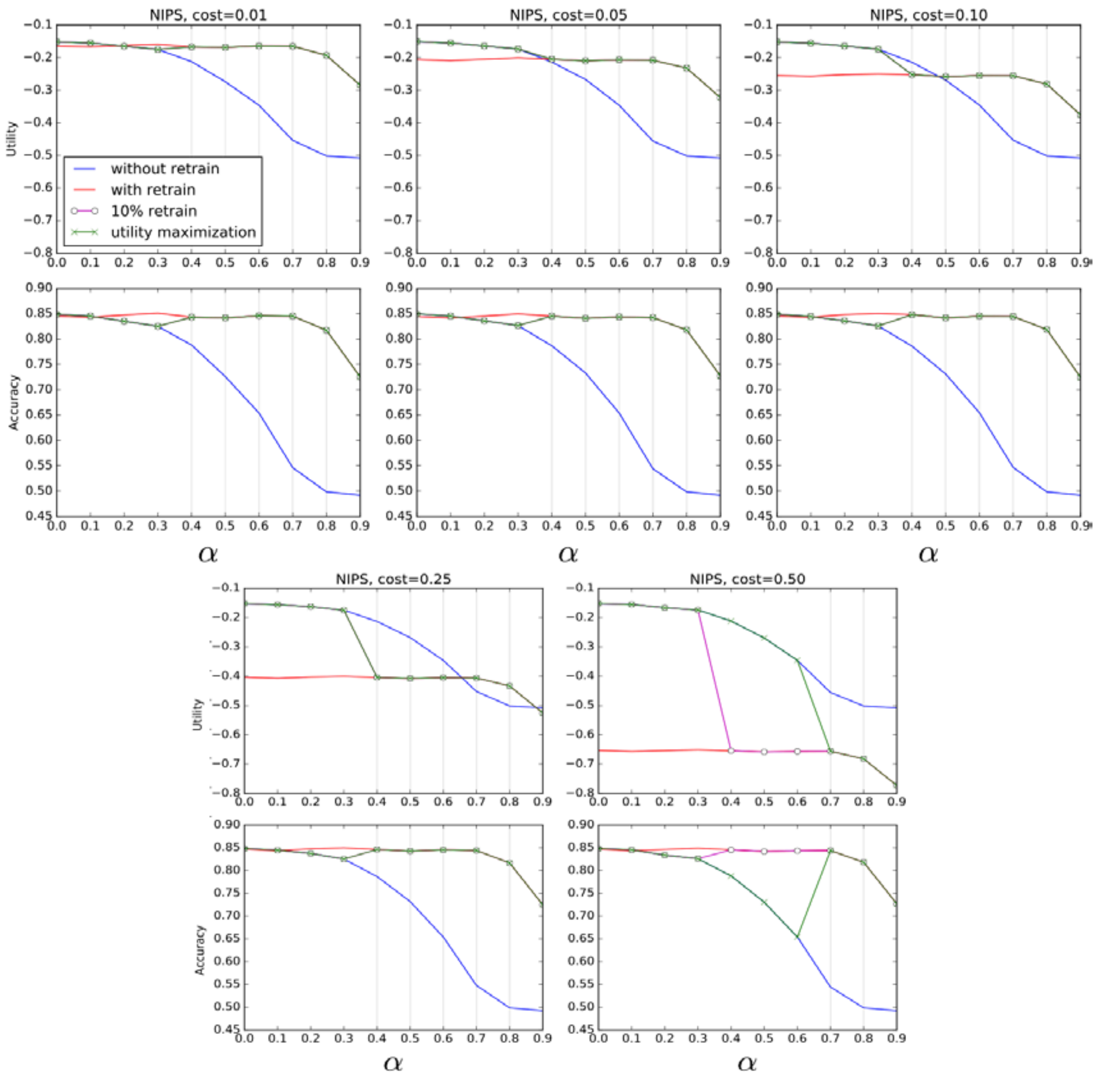


Figure 7 Results for the NIPS dataset. The vertical grey lines indicate "retraining" for our decision-theoretic method

Fig. 7 shows the utility and accuracy as a function of α under the four strategies, and five different values for the cost parameter, C = {0.01, 0.05, 0.1, 0.25, 0.5}.

The results are exactly what we expect: we consistently get a high accuracy of 0.85 if we always retrain, and our accuracy tapers down to 0.5 if we never retrain. The always-retrain strategy achieves high utility when the cost of retraining is low, and the never-retrain strategy achieves high utility when the cost of retraining is high. Both the +10%-surprise and utility-maximization perform about equally well in the low- to mid- retraining cost scenarios, but the +10%-surprise strategy suffers when the cost of retraining is high. Note that by suffering we mean that the utility of the strategy is lower: the accuracy under this strategy is of course better. However, the gains in accuracy are erased by high cost of retraining. Thus, overall, the utility-maximization approach produces better results.

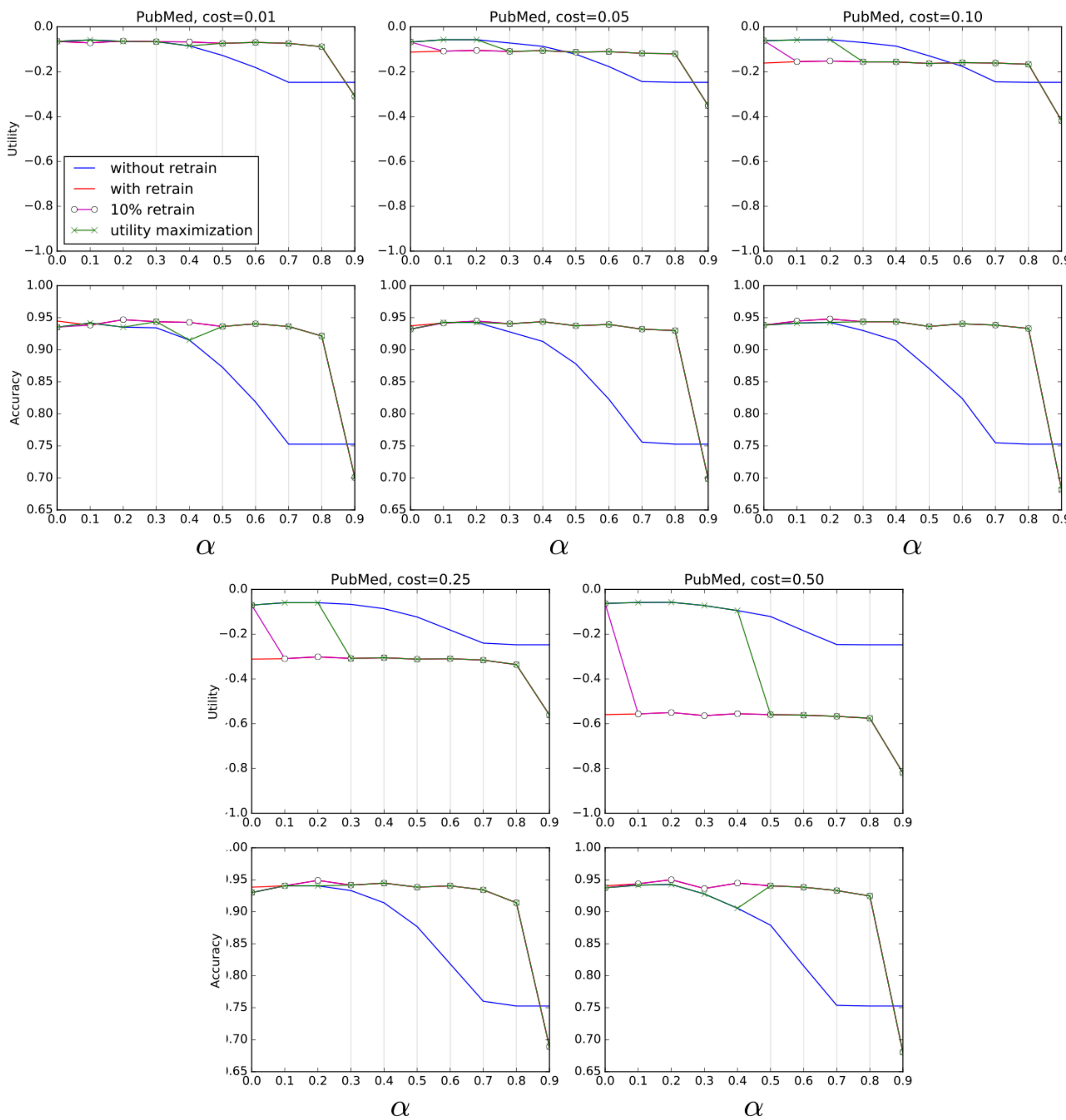


Figure 8 Results for the PubMed dataset. The vertical grey lines indicate "retraining" for our decision-theoretic method.

The results for the PubMed (see Fig. 8) demonstrate the same general behavior. Here, always retraining gets us an accuracy score of about 0.95, and our accuracy dips to 0.75 when we never retrain. The always-retrain strategy edges out the never-retrain strategy when cost is low, but suffers greatly when the cost of retraining is high. The +10%-surprise strategy performs almost no better than the always-retrain strategy; the surprise for this dataset grew rapidly with α, so the +10%-surprise strategy decided to retrain except for very small alpha. We expect this to be the case for at least some datasets, since '+10%' is not a learned constant. The utility-maximization strategy almost always outperforms the +10%-surprise strategy for this dataset. For this dataset especially, the utility-maximization function performs worse than the never-retrain strategy for high values of α. This means a better surprise-to-accuracy estimation function than ours would be less optimistic about retraining when α is large.

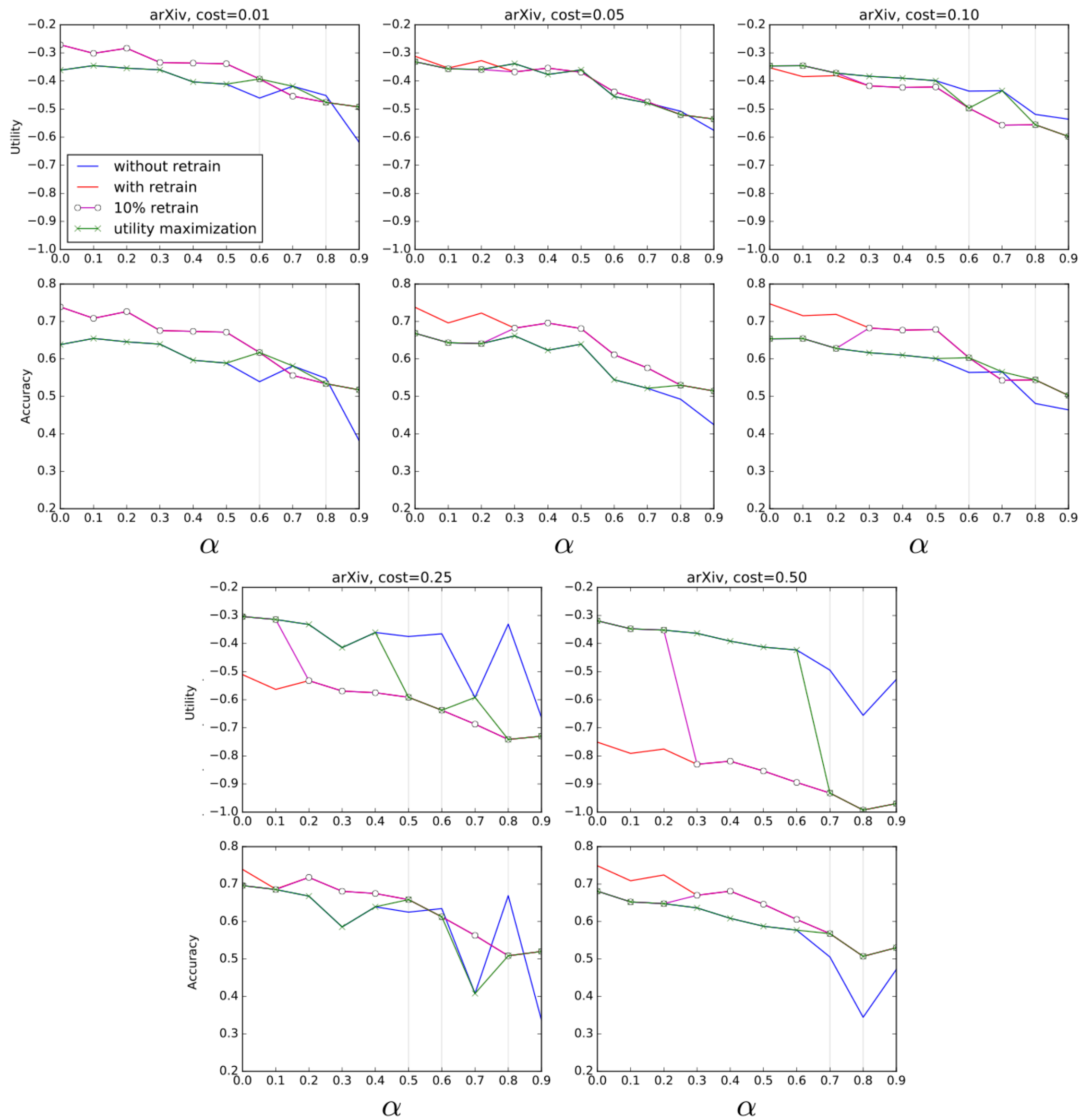


Figure 9 Results for the arxiv dataset. The vertical grey lines indicate "retraining" for our decision-theoretic method

Finally, we focus on the arxiv dataset (Fig. 9). Note that one of the main differences of this dataset from the other two is that the documents are significantly shorter (abstracts instead of full text), thus there are significant fluctuations. For this dataset, retraining does not give a significant improvement in accuracy, so the cost of retraining is the most significant factor in the utility model (although, note that increasing cost does not necessarily mean fewer number of retrainings, due to above mentioned fluctuations). As with the PubMed dataset, the +10%-surprise strategy decides to retrain for all except very small $\boldsymbol{\alpha}$. The performance of the utility-maximization strategy is more mixed here, although, overall, it still yields the most balanced approach to retraining. It sometime performs the best except for when $\boldsymbol{\alpha}$ and the retraining cost are high, in which case the never-retrain strategy performs better. As with the PubMed dataset, our surprise-to-accuracy estimation function should show less affinity to retrain when α is large.

Drift Detection for the Machine Translation Task

Training Machine Translation Engines

Our experiments in the machine translation domain will focus on English-French parallel corpora-based translations. We focused on two main datasets, D1=OpenSubtitles2015 (***os***), which contains subtitles from movies, and D2= MultiUN (***mun***), which is a multilingual corpus from the United Nations documents.

Based on those two corpora, we trained three MT engines, M1, M2, and M3. The first two engines have been trained on D1 and D2, respectively, whereas M3 has been trained on the union of two corpora D1+D2.

We evaluate the quality of the given MT engine (when applied to a given dataset) by the so-called BLEU Score (see https://en.wikipedia.org/wiki/BLEU), which is the adopted metric in the MT research community.

File Name	BLEU(M1)	BLEU(M2)	BLEU(M3)
en/2005/UNEP_POPS_COP1_12.xml.gz	33.9	10.9	33.6
en/2005/A_C5_60_L22.xml.gz	76.8	8.4	76.7
en/2005/CD_PV971.xml.gz	46.2	11.9	44.6
en/2005/FCCC_KP_CMP_2005_6.xml.gz	32.5	13.1	32.5
en/2005/A_C1_60_L33_REV1.xml.gz	69.7	13.9	68.4
en/2005/S_AC45_2005_27.xml.gz	34.8	12.4	32
en/2005/E_CN4_2005_L63.xml.gz	73.6	16.1	73.8
en/2005/TRANS_WP29_2005_82.xml.gz	76.2	10.9	77.7
en/2005/CCPR_C_83_D_823_1998.xml.gz	37.9	12.7	36.5
en/2005/E_2005_L51.xml.gz	55.5	13.3	53.7
en/2005/A_60_PV17.xml.gz	57.5	14.8	55.1
en/2005/HBP_WP7_2005_8.xml.gz	23.6	9.88	22.9
en/2005/S_PV5277.xml.gz	52.2	12.5	50.8
en/2005/E_CN4_SUB2_2005_L40.xml.gz	69.8	21	71.7
en/2005/FCCC_KP_CMP_2005_3.xml.gz	41.2	15.1	41.7
en/2005/S_2005_494.xml.gz	50.2	20.7	52.5
en/2005/NPT_CONF2005_MCIII_WP2.xml.gz	67.0	15.9	70.2

Partial output of the trained MT engines on dataset D1 is shown in the table above. The first column shows the name of the documents (5000 in the test dataset). The second, third and fourth columns show the BLEU scores of models M1, M2, and M3, respectively, for the corresponding document. Note that the BLEU score of M2 (column 2) are considerably smaller than BLEU(M1). This is of course due to the fact that the M2 is trained on a different dataset (D2), and the relatively poor performance is due to domain mismatch between D1 and D2.

Surprise vs Drift

We have examined this phenomenon in a more fine-grained manner, by constructing a test set that was a tunable mixture of D_1 and D_2, $D_{Test} = (1-\alpha)D_1 + \alpha D_2$. Thus, $a = 0$ and $a = 1$ corresponds to no drift and maximum drift, respectively.

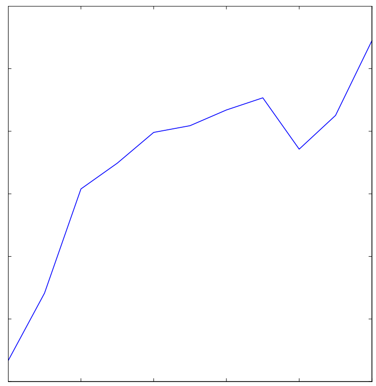


Figure 10 Surprise as a function of mixing parameter alpha

The results are shown in Figure 10. The relationship is mostly what we expect, with surprise increasing with a. One exception is for $a = 0.7$ where the surprise had a slight decrease, but then it starts increasing again. We believe this counterintuitive decrease will disappear if we average the results for many random trials.

Domain Drift and Translation Accuracy

Next, we study the relationship between the amount of domain drift (as measured by surprise) and the translation accuracy as measured by BLEU scores.

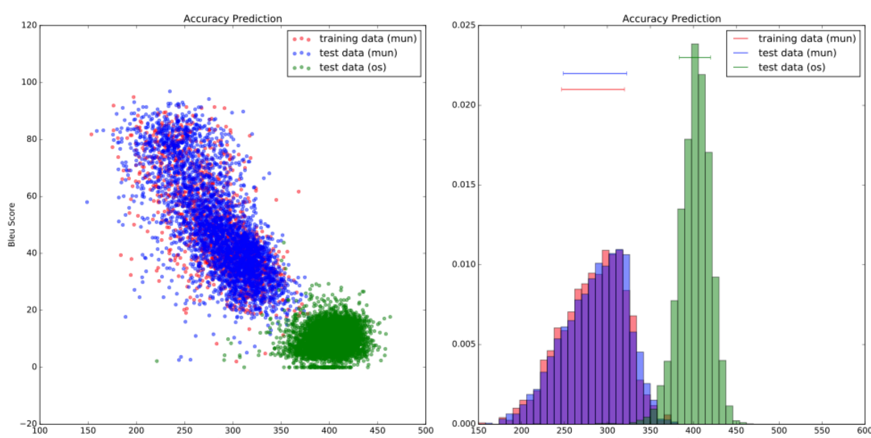


Figure 11 (Left) Scatter plot of BLEU vs Surprise, where each point is a document; training and test sets are as indicated in the legend. (Right) Histogram of Surprise for training and test sets

Figure 11 shows the scatter plot of the BLEU scores vs surprise, when the *mun* is the reference dataset and *os* is the test dataset. There are several worthwhile observations we can make. First, we see that there are two well-separated clusters of documents corresponding to either datasets. Second, when the test set is also chosen from *mun*, there is no discernable differences between the train and test sets; see the figure on the right where we show the histogram of the Surprise for all three datasets. Finally, document level BLEU score is ***decreasing with surprise***, so that more surprising documents are translated less accurately.

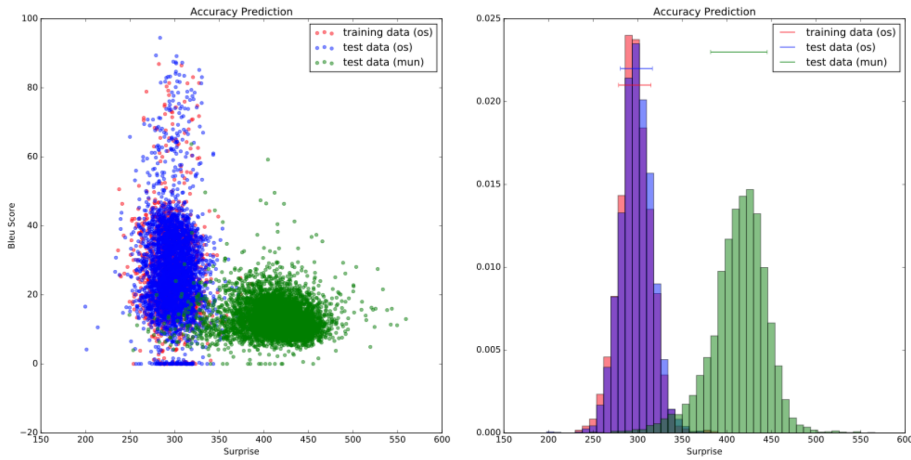


Figure 12 Same as in Figure 2 but for different train and test split; see the legend

Similar picture albeit with some differences is observed when we train on *os* and then test on *mun*. Namely, there are still two well-separated clusters. However, in this case, the relationship between the BLEU scores and Surprise in the training dataset is much more random. Namely, two documents might have the same surprise, but their BLEU scores can different by significant amount. Note also that there are some documents in the test set (mun) that have higher BLEU score than some of the documents in the training set, even if those documents have higher values of surprise. We are planning to analyze this phenomenon in more details in coming weeks.

<u>Toward Active Drift Correction Methods</u>

We have also conducted experiments with more elaborate retraining cost models compared to what we had considered for the topic modeling problem. Remarkably, this type of cost models are omnipresent in MT domain. Namely, given two domains such as mun and os, and the distributional mismatch as measured by Surprise, we can ask the following questions:

1. If we are getting higher surprise in the test dataset, how much we will gain if we spend some budget on annotating additional data (for MT, annotating means manually building a parallel corpus)?
2. For a given budget, which of the documents one should translate for building that parallel corpus?

For the second question, the baseline approach would be to select documents at random. However, another intuitive approach would be selecting the documents based on their Surprise, e.g., documents that have higher surprise should get higher priority for annotation.

A full analysis of the above strategy would correspond to training more MT engines with different sets of parallel corpora, which is a very costly exercise, and given the limited time we have for the program, might not be feasible. Instead, we conducted an alternative set of experiments, where, instead of evaluating the data selection approach on translation accuracy, we evaluate it based on how much it reduces surprise.

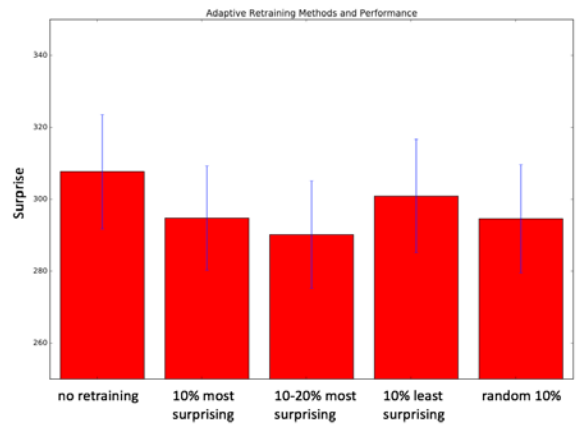


Figure 13 Surprise under different data selection strategies

The results are shown in Figure 13. First, we rank all the documents in the test set according to the Surprise, e.g., top 10%, 10-20%, …, bottom 10%. In addition to the baseline method with no retraining, we consider 4 different data selection strategies: (1) Select from the top 10%; (2) Select from 10%-20%; (3) Select from the bottom 10%; (4) Select randomly. Under all four strategies we observe decrease in surprise, which is intuitive. Furthermore, the decrease is the weakest under strategy (3), which is also understandable, since the documents that are not so surprising were already well-represented in the original training set, and including them again will not change much. Perhaps the more interesting findings are that selecting the top 10% results in the same decrease in surprise as selecting randomly, and that selecting from 10%-20% yields the best reduction in surprise. This is probably because this range of surprise includes documents that are typical, and not just outliers in the test set. However, this point needs further examination.

Conclusions

To conclude, we have proposed a novel computational framework for detecting and quantifying model drift, and correcting drift based on decision-theoretic framework. We have also performed exhaustive experiments for validating and evaluating the proposed framework. In our first evaluation, the experiments for drift detection and quantification confirmed that surprise as measured by CorEx is indeed able to capture important distributional changes. Furthermore, our experiments also helped with understanding the relationship between drift and performance deterioration. While our results for temporal/gradual drift are not very conclusive, for the abrupt drift scenario we find that there is significant statistical relationship between increase in surprise and performance deterioration. Importantly, the relationship seems to be qualitatively similar for different datasets (albeit with quantitative difference that are expected).

In the second evaluation, we found that our proposed decision-theoretic drift-correction framework performed as expected. Specifically, the advantage of the proposed approach is its ability to adapt to different cost/benefit ratio of a given scenario. Indeed, for low cost of retraining, the behavior produced by the utility-maximization approach is similar to "always retrain" and "10% retrain" strategies, while for larger C, it starts to become more similar to "never retrain" strategy. This adaptive nature of the proposed method makes it the best overall choice among the baselines, when the performance is measured via the utility function.

Recommendations

Based on the findings of our project, we believe there are several important directions where further explorations are needed. First of all, one of the central problems we encountered within our seedling project was the performance prediction, e.g., ability to predict the performance of an algorithm trained on one dataset, when that algorithm is used on a previously unseen dataset. While this is an active research area for domains such as machine translation, we believe that efficient solutions to this problem can be relevant and valuable for diverse set of machine learning applications. On a more general note, while our project has addressed specific aspects of model drift phenomenon, we believe there is a need for a more general and broader research agenda for machine learning in time-varying and non-stationary environments.

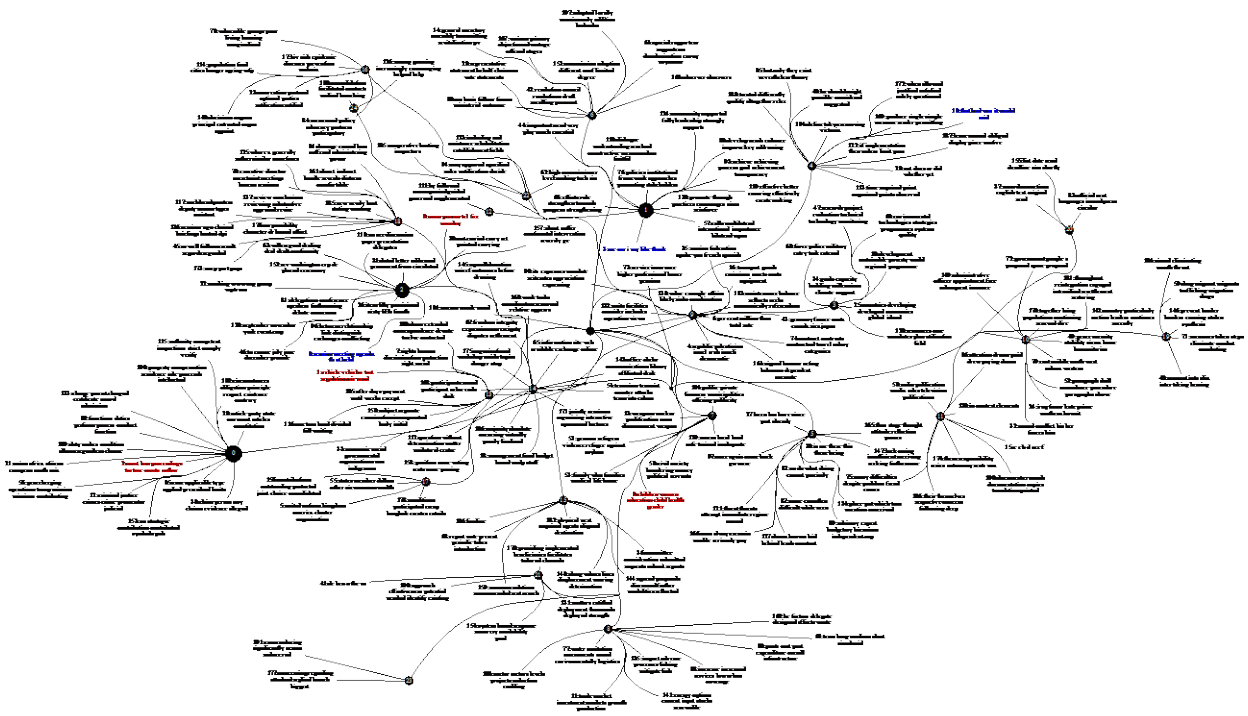
References

[1] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In Proc. of NIPS'14, 2014.

[2] Greg Ver Steeg and Aram Galstyan. Maximally Informative Hierarchical Representations of High-Dimensional Data. In Proc. of AISTATS'15, 2015.

Appendix

A. Hierarchical structure learned by Corex on mun dataset



B. Topics learned by Corex on mun dataset

Below we provide the list of topics discovered by CorEx for the MultiUN dataset. There are two line for each topic: The first line shows the Group number corresponding to a latent variable, and the total correlation TC(X;Y_j) between that latent variable and words in that topic. The second line shows the top words that are most relevant to that topic.

When running CorEx, the number of latent variables (and hence # of topics) was set to 200.

Group num: 0, TC(X;Y_j): 0.690
0:children,women,education,child,health,gender,school,care,age,men
Group num: 1, TC(X;Y_j): 0.546
1:vehicle,vehicles,test,regulation,air,used,mm,manufacturer,amend,temperature
Group num: 2, TC(X;Y_j): 0.489
2:court,law,proceedings,torture,courts,author,act,detention,cases,offence
Group num: 3, TC(X;Y_j): 0.432
3:we,our,i,my,like,thank,us,me,hope,today
Group num: 4, TC(X;Y_j): 0.407
4:republic,palestinian,israel,arab,israeli,democratic,congo,mr,president,occupied
Group num: 5, TC(X;Y_j): 0.383
5:united,nations,kingdom,america,charter,organization,bretton,woods,summits,acco rding
Group num: 6, TC(X;Y_j): 0.344
6:per,cent,million,than,total,rate,estimated,average,years,less
Group num: 7, TC(X;Y_j): 0.340
7:rights,human,discrimination,protection,right,racial,cultural,freedoms,fundamental ,promotion
Group num: 8, TC(X;Y_j): 0.327
8:room,pm,am,tel,fax,monday,mail,wednesday,thursday,friday
Group num: 9, TC(X;Y_j): 0.272
9:session,meeting,agenda,th,at,held,hoc,ad,seventh,twenty
Group num: 10, TC(X;Y_j): 0.257
10:that,had,was,it,would,said,were,noted,could,stated
Group num: 11, TC(X;Y_j): 0.257
11:trade,market,investment,markets,growth,production,economy,agricultural,prod ucts,business
Group num: 12, TC(X;Y_j): 0.246
12:criminal,justice,crimes,crime,prosecutor,judicial,judges,prison,acts,prosecution
Group num: 13, TC(X;Y_j): 0.233
13:weapons,nuclear,proliferation,arms,disarmament,weapon,destruction,treaty,iaea ,npt
Group num: 14, TC(X;Y_j): 0.227

14:general,secretary,assembly,transmitting,revitalization,pv,heads,moon,fullest,personalities
Group num: 15, TC(X;Y_j): 0.222
15:c,e,b,d,see,f,ii,annex,cn,para
Group num: 16, TC(X;Y_j): 0.194
16:transport,goods,emissions,assets,costs,equipment,carriage,transactions,creditor,creditors
Group num: 17, TC(X;Y_j): 0.175
17:hiv,aids,epidemic,diseases,prevention,malaria,infection,disease,unaids,tuberculosis
Group num: 18, TC(X;Y_j): 0.175
18:management,fund,budget,board,undp,staff,financial,funds,activities,funding
Group num: 19, TC(X;Y_j): 0.167
19:development,sustainable,poverty,world,regional,programme,environment,summit,cooperation,eradication
Group num: 20, TC(X;Y_j): 0.153
20:article,party,state,covenant,articles,constitution,provisions,code,under,art
Group num: 21, TC(X;Y_j): 0.151
21:union,africa,african,european,south,asia,caribbean,latin,pacific,region
Group num: 22, TC(X;Y_j): 0.147
22:working,www,org,group,wp,trans,http,informal,htm,ended
Group num: 23, TC(X;Y_j): 0.145
23:convention,protocol,optional,parties,ratification,ratified,conventions,protocols,treaties,instruments
Group num: 24, TC(X;Y_j): 0.140
24:iraq,timor,leste,prime,northern,kuwait,iraqi,kosovo,minister,sri
Group num: 25, TC(X;Y_j): 0.127
25:countries,developing,developed,economies,global,island,small,least,transition,landlocked
Group num: 26, TC(X;Y_j): 0.121
26:item,fifty,provisional,sixty,fifth,fourth,second,ninth,third,forty
Group num: 27, TC(X;Y_j): 0.117
27:been,has,have,since,past,already,several,begun,completed,gone
Group num: 28, TC(X;Y_j): 0.116
28:not,does,or,did,whether,yet,nor,either,neither,necessarily
Group num: 29, TC(X;Y_j): 0.111
29:representative,statement,behalf,chairman,vote,statements,representatives,vice,election,elected
Group num: 30, TC(X;Y_j): 0.100
30:out,carried,carry,set,pointed,carrying,sets,carries,pointing,setting
Group num: 31, TC(X;Y_j): 0.100
31:dated,letter,addressed,permanent,from,circulated,letters,verbale,herewith,identical
Group num: 32, TC(X;Y_j): 0.100
32:armed,conflict,his,her,forces,him,displaced,civilians,conflicts,war
Group num: 33, TC(X;Y_j): 0.098

33:economic,social,governmental,organizations,non,indigenous,socio,peoples,institutions,participation
Group num: 34, TC(X;Y_j): 0.097
34:goals,capacity,building,millennium,climate,support,change,lessons,partnerships,learned
Group num: 35, TC(X;Y_j): 0.094
35:russian,federation,spoke,you,french,spanish,arabic,your,chinese,sir
Group num: 36, TC(X;Y_j): 0.094
36:committee,consideration,submitted,requests,submit,reports,recommends,notes,observations,requested
Group num: 37, TC(X;Y_j): 0.093
37:record,corrections,english,text,original,read,copy,insert,verbatim,rose
Group num: 38, TC(X;Y_j): 0.089
38:claim,person,any,claims,evidence,alleged,claimant,facts,finds,panel
Group num: 39, TC(X;Y_j): 0.088
39:is,are,there,this,these,being,however,still,even,most
Group num: 40, TC(X;Y_j): 0.087
40:peace,security,stability,sierra,leone,humanitarian,sudan,darfur,afghanistan,lasting
Group num: 41, TC(X;Y_j): 0.087
41:de,la,n,o,the,m,facto,et,des,of
Group num: 42, TC(X;Y_j): 0.083
42:resolution,council,resolutions,draft,recalling,pursuant,reaffirming,sponsors,declaration,res
Group num: 43, TC(X;Y_j): 0.081
43:germany,france,costa,canada,rica,japan,italy,netherlands,australia,norway
Group num: 44, TC(X;Y_j): 0.074
44:important,need,very,play,much,essential,success,strong,crucial,good
Group num: 45, TC(X;Y_j): 0.072
45:as,well,follows,result,regards,regarded,serve,whole,insofar,viewed
Group num: 46, TC(X;Y_j): 0.071
46:to,ensure,july,june,december,provide,march,necessary,october,april
Group num: 47, TC(X;Y_j): 0.071
47:research,project,evaluation,technical,technology,monitoring,institute,science,studies,analysis
Group num: 48, TC(X;Y_j): 0.069
48:account,into,alia,inter,taking,bearing,mind,take,incorporation,chase
Group num: 49, TC(X;Y_j): 0.065
49:be,should,might,possible,considered,suggested,given,soon,desirable,acceptable
Group num: 50, TC(X;Y_j): 0.063
50:drug,migrant,migrants,trafficking,migration,drugs,workers,narcotic,smuggling,undcp
Group num: 51, TC(X;Y_j): 0.062
51:persons,refugees,violence,refugee,against,asylum,disabilities,victims,unhcr,camps
Group num: 52, TC(X;Y_j): 0.061

52:paragraph,shall,accordance,procedure,paragraphs,above,rule,referred,described,subparagraph
Group num: 53, TC(X;Y_j): 0.060
53:family,who,families,medical,life,home,woman,hospital,psychological,live
Group num: 54, TC(X;Y_j): 0.058
54:terrorism,terrorist,counter,attacks,terrorists,cuban,suppression,cuba,taliban,qaida
Group num: 55, TC(X;Y_j): 0.056
55:states,member,dollars,other,oic,commonwealth,sovereign,mutual,bush,participating
Group num: 56, TC(X;Y_j): 0.053
56:peacekeeping,operations,troop,mission,missions,contributing,contributors,monuc,unamsil,stabilization
Group num: 57, TC(X;Y_j): 0.052
57:calls,multilateral,international,importance,bilateral,upon,agreements,commitment,reaffirms,continue
Group num: 58, TC(X;Y_j): 0.052
58:radio,publication,media,sales,television,publications,published,broadcasting,print,broadcast
Group num: 59, TC(X;Y_j): 0.050
59:civil,society,laundering,money,political,servants,aviation,servant,makeup,fortune
Group num: 60, TC(X;Y_j): 0.049
60:force,police,military,entry,task,entered,civilian,officers,personnel,enter
Group num: 61, TC(X;Y_j): 0.048
61:term,long,medium,short,sized,mid,beginning,remainder,haul,nigger
Group num: 62, TC(X;Y_j): 0.046
62:high,commissioner,level,ranking,tech,sin,leonard,wan,bump,jam
Group num: 63, TC(X;Y_j): 0.045
63:with,regard,dealing,deal,dealt,conformity,connection,line,associated,conjunction
Group num: 64, TC(X;Y_j): 0.043
64:special,rapporteur,rapporteurs,decolonization,envoy,myanmar,visit,colonialism,visits,visiting
Group num: 65, TC(X;Y_j): 0.039
65:information,site,web,available,exchange,online,sites,readily,accessible,disseminate
Group num: 66, TC(X;Y_j): 0.037
66:efforts,role,strengthen,towards,progress,strengthening,played,comprehensive,reform,implement
Group num: 67, TC(X;Y_j): 0.036
67:freedom,integrity,expression,sovereignty,disputes,settlement,territorial,dispute,independence,belief
Group num: 68, TC(X;Y_j): 0.033
68:report,note,present,periodic,takes,introduction,questions,detailed,hrc,endorses
Group num: 69, TC(X;Y_j): 0.032
69:environmental,technologies,strategies,programmes,systems,quality,knowledge,indicators,tools,frameworks

Group num: 70, TC(X;Y_j): 0.031
70:east,middle,north,west,sahara,western,near,atlantic,hills,lawn
Group num: 71, TC(X;Y_j): 0.029
71:measures,taken,steps,eliminate,combat,combating,preventive,corruption,preventing,anti
Group num: 72, TC(X;Y_j): 0.026
72:government,people,s,proposed,space,proposal,outer,additional,foreign,uses
Group num: 73, TC(X;Y_j): 0.026
73:service,insurance,higher,professional,lower,pension,employees,providers,fees,career
Group num: 74, TC(X;Y_j): 0.025
74:contract,contracts,contractual,travel,salary,categories,allowance,salaries,category,temporary
Group num: 75, TC(X;Y_j): 0.024
75:many,difficulties,despite,problem,faced,causes,face,remains,recent,decades
Group num: 76, TC(X;Y_j): 0.024
76:policies,institutional,framework,approaches,promoting,stakeholders,issues,initiatives,improving,mechanisms
Group num: 77, TC(X;Y_j): 0.022
77:water,sanitation,assessments,sound,environmentally,logistics,base,electricity,drinking,assessment
Group num: 78, TC(X;Y_j): 0.022
78:executive,director,secretariat,meetings,bureau,sessions,administrator,consultation,preparation,steering
Group num: 79, TC(X;Y_j): 0.019
79:vulnerable,groups,poor,living,housing,marginalized,affected,increasing,socially,unemployment
Group num: 80, TC(X;Y_j): 0.019
80:posts,cost,post,expenditure,overall,infrastructure,expected,operational,external,savings
Group num: 81, TC(X;Y_j): 0.018
81:delegations,conference,speakers,forthcoming,debate,consensus,convening,discussions,advance,intend
Group num: 82, TC(X;Y_j): 0.017
82:so,do,what,doing,cannot,precisely,lose,afford,sight,reason
Group num: 83, TC(X;Y_j): 0.017
83:official,sent,languages,issued,press,circular,interpreters,gazette,written,received
Group num: 84, TC(X;Y_j): 0.016
84:areas,rural,policy,advocacy,partners,participatory,capacities,makers,decentralization,integration
Group num: 85, TC(X;Y_j): 0.016
85:but,only,they,exist,nevertheless,theory,properly,confined,reversed,picking
Group num: 86, TC(X;Y_j): 0.015
86:attention,drawn,paid,drew,paying,draws,amazing
Group num: 87, TC(X;Y_j): 0.015
87:some,can,often,difficult,while,seen,way,both,become,far

Group num: 88, TC(X;Y_j): 0.015
88:on,basis,follow,forum,ministerial,outcome,conferences,intergovernmental,thematic,concentrate
Group num: 89, TC(X;Y_j): 0.015
89:advisory,expert,budgetary,biennium,independent,cop,experts,cp,biennial,unfccc
Group num: 90, TC(X;Y_j): 0.014
90:its,expresses,mandate,reiterates,appreciation,expressing,endorsed,reiterated,expeditiously,literature
Group num: 91, TC(X;Y_j): 0.014
91:damage,caused,loss,suffered,administering,power,causing,cause,lost,compensate
Group num: 92, TC(X;Y_j): 0.014
92:functions,duties,perform,powers,conduct,function,responsible,statutory,performing,confidentiality
Group num: 93, TC(X;Y_j): 0.014
93:achieve,achieving,process,goal,achievement,transparency,accountability,contribute,transparent,achieved
Group num: 94, TC(X;Y_j): 0.013
94:may,approval,specified,rules,notification,decide,prior,reference,listed,receipt
Group num: 95, TC(X;Y_j): 0.011
95:case,applicable,type,applied,prescribed,limits,applies,defined,specify,partial
Group num: 96, TC(X;Y_j): 0.011
96:between,relationship,link,distinguish,exchanges,conflicting,devil,derek,tooth,crying
Group num: 97, TC(X;Y_j): 0.011
97:once,again,come,back,go,mere,never,thing,tell,says
Group num: 98, TC(X;Y_j): 0.010
98:increase,increased,services,low,urban,coverage,skills,remote,volunteers,generating
Group num: 99, TC(X;Y_j): 0.010
99:develop,needs,enhance,improve,key,addressing,assist,objectives,facilitate,strengthened
Group num: 100, TC(X;Y_j): 0.010
100:documents,records,documentation,copies,translation,printed,page,pages,certified,versions
Group num: 101, TC(X;Y_j): 0.010
101:cross,reducing,significantly,across,reduce,red,gap,cutting,greater,pace
Group num: 102, TC(X;Y_j): 0.009
102:circumstances,obligation,principle,respect,existence,contrary,considers,distinction,constitute,accept
Group num: 103, TC(X;Y_j): 0.009
103:sector,sectors,levels,projects,reduction,enabling,structural,grants,incentives,learning
Group num: 104, TC(X;Y_j): 0.009
104:public,private,finances,municipalities,offering,publicity,branches,besides,treasure
Group num: 105, TC(X;Y_j): 0.008

105:after,days,payment,until,weeks,except,exceeding,exceed,termination,suspended
Group num: 106, TC(X;Y_j): 0.008
106:their,themselves,respective,concern,following,deep,especially,approved,others,recognized
Group num: 107, TC(X;Y_j): 0.008
107:adopted,l,orally,unanimously,addition,barbados,frank
Group num: 108, TC(X;Y_j): 0.008
108:property,compensation,residence,sale,proceeds,intellectual,permits,ownership,restitution,deed
Group num: 109, TC(X;Y_j): 0.008
109:approach,effectiveness,potential,needed,identify,existing,priority,identified,assess,identifying
Group num: 110, TC(X;Y_j): 0.008
110:dialogue,understanding,reached,constructive,memorandum,fruitful,participate,restricted,tripartite,unknown
Group num: 111, TC(X;Y_j): 0.006
111:by,followed,accompanied,guided,governed,supplemented,backed,thereafter,envelope
Group num: 112, TC(X;Y_j): 0.006
112:question,without,determination,matter,unilateral,centre,prejudice,proceed,answer,resolved
Group num: 113, TC(X;Y_j): 0.006
113:time,required,point,organized,points,observed,delays,frame,terms,uncertainty
Group num: 114, TC(X;Y_j): 0.006
114:population,food,cities,hunger,ageing,wfp,madrid,launched,bridge,repercussions
Group num: 115, TC(X;Y_j): 0.006
115:authority,competent,inspections,strict,comply,verify,complying,purposes,seabed,discovery
Group num: 116, TC(X;Y_j): 0.006
116:one,two,hand,divided,fall,waiting,expense,rob,writes,fame
Group num: 117, TC(X;Y_j): 0.006
117:such,headquarters,deputy,means,types,assistant,adviser,coordinator,nature,liaison
Group num: 118, TC(X;Y_j): 0.005
118:un,ece,discussion,paper,presentation,delegates,subsidiary,cefact,ensuing,doc
Group num: 119, TC(X;Y_j): 0.005
119:effective,better,ensuring,effectively,create,making,best,creating,objective,encourage
Group num: 120, TC(X;Y_j): 0.005
120:access,local,land,safe,trained,inadequate,provinces,districts,aid,councils
Group num: 121, TC(X;Y_j): 0.005
121:threat,threats,attempt,immediate,regime,annual,commit,threatened,pose,refrain
Group num: 122, TC(X;Y_j): 0.005
122:if,implementation,then,unless,limit,pass,escape,exact,discovered,sit
Group num: 123, TC(X;Y_j): 0.005

123:including,and,assistance,rehabilitation,establishment,fields,withdrawn
Group num: 124, TC(X;Y_j): 0.004
124:community,supported,fully,leadership,strongly,supports,called,renewed,urgently,pillar
Group num: 125, TC(X;Y_j): 0.004
125:where,a,generally,rather,similar,sometimes,longer,difference,consequence,becomes
Group num: 126, TC(X;Y_j): 0.004
126:among,growing,increasingly,encouraging,helped,help,helping,active,things,friendly
Group num: 127, TC(X;Y_j): 0.004
127:down,known,laid,behind,leads,constant,run,exit,allowing,fairly
Group num: 128, TC(X;Y_j): 0.003
128:resources,core,mandates,plan,utilization,field,enhancement,utilize,genetic,undertaken
Group num: 129, TC(X;Y_j): 0.003
129:september,november,york,event,cmp
Group num: 130, TC(X;Y_j): 0.003
130:promote,through,practices,encourages,aims,reinforce,met,complemented
Group num: 131, TC(X;Y_j): 0.003
131:matters,entitled,deployment,thousands,deployed,strength,status,start,direction,driven
Group num: 132, TC(X;Y_j): 0.003
132:units,facilities,consider,includes,operation,views,made,activity,formed,owned
Group num: 133, TC(X;Y_j): 0.003
133:charge,parent,charged,certificate,award,admission,german,admitted,awarded,awards
Group num: 134, TC(X;Y_j): 0.003
134:place,put,which,turn,mention,conceived,assumed
Group num: 135, TC(X;Y_j): 0.003
135:impact,adverse,processes,fishing,mitigate,fish,migratory,capabilities,catch,complement
Group num: 136, TC(X;Y_j): 0.002
136:seminar,ngo,chaired,briefings,hosted,dpi,symposium,ababa,addis,fellowship
Group num: 137, TC(X;Y_j): 0.002
137:review,conclusions,reviewing,substantive,appraisal,revise,thorough,severe,isolated,fabric
Group num: 138, TC(X;Y_j): 0.002
138:in,context,elements
Group num: 139, TC(X;Y_j): 0.002
139:value,example,affairs,likely,risks,combination,depends,easily,flexible,real
Group num: 140, TC(X;Y_j): 0.002
140:administrative,officer,appointment,free,subsequent,issuance,appointments,preliminary,branch,zones
Group num: 141, TC(X;Y_j): 0.002

141:energy,options,current,input,stocks,renewable,meet,reply,aforementioned,geog
raphical
Group num: 142, TC(X;Y_j): 0.002
142:country,particularly,section,leaders,continues,recently,bringing,notably,pursue
d,ties
Group num: 143, TC(X;Y_j): 0.002
143:office,ohchr,communications,library,affiliated,desk,center,advisor,fresh
Group num: 144, TC(X;Y_j): 0.002
144:agreed,proposals,discussed,further,modalities,reflected,implications,outlined,re
gistered,immigration
Group num: 145, TC(X;Y_j): 0.002
145:up,collaboration,unicef,outcomes,before,drawing,exception,foundation,explanat
ion,clusters
Group num: 146, TC(X;Y_j): 0.002
146:prevent,border,borders,crossing,stolen,synthesis,recovering,pep
Group num: 147, TC(X;Y_j): 0.002
147:lack,owing,insufficient,receiving,seeking,furthermore,lacking,formal,sought,pro
blematic
Group num: 148, TC(X;Y_j): 0.002
148:along,values,lines,displacement,moving,deterioration,governing,steady,shape,p
ressures
Group num: 149, TC(X;Y_j): 0.002
149:decisions,organs,principal,entrusted,organ,appoint,demonstration,tend,chain,c
oupled
Group num: 150, TC(X;Y_j): 0.002
150:subject,separate,examination,incorporated,body,initial,covered,examined,settle
ments,mentioned
Group num: 151, TC(X;Y_j): 0.002
151:commission,adoption,different,conf,limited,degree,operate,multiple,routine,ben
eficiary
Group num: 152, TC(X;Y_j): 0.002
152:rev,washington,crp,dc,placed,ceremony,requires,ensured,forming,tom
Group num: 153, TC(X;Y_j): 0.001
153:no,strategic,contribution,contributed,symbols,pub,ya
Group num: 154, TC(X;Y_j): 0.001
154:system,based,response,recovery,availability,pool,observing
Group num: 155, TC(X;Y_j): 0.001
155:list,date,send,deadline,aim,shortly,advised,nominated,postponed,sphere
Group num: 156, TC(X;Y_j): 0.001
156:contributions,outstanding,protected,joint,choice,consolidated,exclusive,acquire,
abandoned,belong
Group num: 157, TC(X;Y_j): 0.001
157:about,suffer,continental,intervention,severely,gc,kinds,shelf,every,disproportio
nate
Group num: 158, TC(X;Y_j): 0.001
158:position,same,voting,seats,none,passing,reserved,having,yes,discharged

Group num: 159, TC(X;Y_j): 0.001
159:recommendations,recommended,rest,search
Group num: 160, TC(X;Y_j): 0.001
160:work,tasks,coordinators,removed,relative,appears,upcoming,settled,noticed
Group num: 161, TC(X;Y_j): 0.001
161:direct,indirect,handle,reveals,distress,comfortable,turns
Group num: 162, TC(X;Y_j): 0.001
162:he,factors,delegate,designed,effects,wrote,suited,adapted,samuel,incorporating
Group num: 163, TC(X;Y_j): 0.001
163:observer,observers
Group num: 164, TC(X;Y_j): 0.001
164:signed,honour,acting,bahamas,dependent,accurate,sensitivity,anthony,deficienc
ies,phillip
Group num: 165, TC(X;Y_j): 0.001
165:thus,stage,thought,attitude,reflection,proves,anywhere,mistaken
Group num: 166, TC(X;Y_j): 0.001
166:own,always,remain,unable,seriously,pay,equally,assume,giving,trying
Group num: 167, TC(X;Y_j): 0.001
167:various,primary,objection,advantage,offered,stages,created,sensitive,linked,rela
tionships
Group num: 168, TC(X;Y_j): 0.001
168:participants,round,participant,eclac,ends,dark
Group num: 169, TC(X;Y_j): 0.001
169:produce,single,simple,measure,render,permitting,typical,replacing,leaves,insta
nt
Group num: 170, TC(X;Y_j): 0.001
170:together,bring,populations,continuing,renewal,dire,willingness,goose
Group num: 171, TC(X;Y_j): 0.001
171:jointly,seminars,organizing,interactive,sponsored,lectures,intact
Group num: 172, TC(X;Y_j): 0.001
172:when,allowed,justified,satisfied,solely,questioned,exactly,aside,thoroughly,entir
ely
Group num: 173, TC(X;Y_j): 0.001
173:unep,part,pops
Group num: 174, TC(X;Y_j): 0.001
174:them,responsibility,series,autonomy,rests,usa,summarized,realm,cos
Group num: 175, TC(X;Y_j): 0.000
175:organizational,workshop,unido,topics,danger,stop,consultant,idb,committees,di
sturbed
Group num: 176, TC(X;Y_j): 0.000
176:co,possibility,character,dr,bound,affect,bear,advisers,explicit,proper
Group num: 177, TC(X;Y_j): 0.000
177:concerning,regarding,attached,replied,launch,biggest
Group num: 178, TC(X;Y_j): 0.000
178:conditions,participated,escap,bangkok,creates,entails,star
Group num: 179, TC(X;Y_j): 0.000

179:providing,implemented,beneficiaries,facilitates,tailored,channels
Group num: 180, TC(X;Y_j): 0.000
180:duty,makes,condition,allows,regardless,choose,irrespective,chosen,govern,wee kend
Group num: 181, TC(X;Y_j): 0.000
181:throughout,reintegration,engaged,intensified,resettlement,restoring,demonstra ting,tactics
Group num: 182, TC(X;Y_j): 0.000
182:physical,next,acquired,agents,aligned,destination,documented,timetable,occurr ence,recognise
Group num: 183, TC(X;Y_j): 0.000
183:maintenance,balance,reflects,seeks,economically,referendum,reliance,saving,so phisticated,assumptions
Group num: 184, TC(X;Y_j): 0.000
184:aimed,eliminating,month,thrust
Group num: 185, TC(X;Y_j): 0.000
185:new,newly,host,stating,wasting
Group num: 186, TC(X;Y_j): 0.000
186:majority,absolute,occurring,virtually,poorly,finalized,tough,string
Group num: 187, TC(X;Y_j): 0.000
187:leave,normal,obliged,display,piece,motive
Group num: 188, TC(X;Y_j): 0.000
188:treated,differently,qualify,altogether,relax
Group num: 189, TC(X;Y_j): 0.000
189:draw,extended,correspondence,devote,twelve,contacted,photographs,sixteen,p hotograph,courtesy
Group num: 190, TC(X;Y_j): 0.000
190:consolidation,facilitated,contacts,unified,launching
Group num: 191, TC(X;Y_j): 0.000
191:unesco,unodc,usual
Group num: 192, TC(X;Y_j): 0.000
192:notwithstanding,instances
Group num: 193, TC(X;Y_j): 0.000
193:times,falling,pressed
Group num: 194, TC(X;Y_j): 0.000
194:define,tokyo,removing,victoria
Group num: 195, TC(X;Y_j): 0.000
195:cooperative,hosting,inspectors
Group num: 196, TC(X;Y_j): -0.000
196:finalize
Group num: 197, TC(X;Y_j): -0.000
197:
Group num: 198, TC(X;Y_j): -0.000
198:
Group num: 199, TC(X;Y_j): -0.000
199:inspectors,falls,sunset,hosting

Symbols, Abbreviations, and Acronyms

S – Surprise
TC – Total Correlation
$\boldsymbol{D_{Ref}}$ - Reference dataset
$\boldsymbol{D_{Test}}$ - Test dataset
α - mixing parameter
CorEx – Correlation Explanation
OS - OpenSubtitles2015 dataset
Mun - MultiUN dataset
MT - Machine Translation