IJICTR International Journal of Information & Communication Technology Research ITRC

Volume 7- Number 3- Summer 2015 (63-72)

Density-Based K-Nearest Neighbor Active Learning for Improving Farsi-English Statistical Machine Translation System

Somayeh Bakhshaei
Department of Computer Engineering and Information Technology
Amirkabir University of technology
Tehran, Iran
Bakhshaei@aut.ac.ir

Reza Safabakhsh
Department of Computer Engineering and Information Technology
Amirkabir University of technology
Tehran, Iran
Safa@aut.ac.ir

Shahram Khadivi
Department of Computer Engineering and Information Technology
Amirkabir University of technology
Tehran, Iran
Khadivi@aut.ac.ir

Received: March 16, 2014- Accepted: May 27, 2015

Abstract—**Labeled data are useful resources for different application in different fields like image processing, natural language processing etc. Producing labeled data is a costly process. One efficient solution for alleviating the costly process of annotating data is managing the sampling process. It is better to query for essential samples instead of a group of unnecessary ones. Active learning (AL) attempts to overcome the labeling bottleneck by sending queries for unlabeled instances to be labeled with the help of an annotator. This technique is applied to Natural Language Processing (NLP) especially in Statistical Machine Translation (SMT) tasks that we also focus on in this work. In Statistical Machine Translation, parallel corpora are scarce resources, and AL is a way of solving this problem. It attempts to alleviate the costly process of data annotating by sending queries just for translation of the most informative sentences which are essential for system improvement. The contribution of our work is proposing a new approach in AL for selecting sentences through a soft decision making process. In this algorithm, in addition to scoring sentences according to their information, the distribution of the space of unlabeled data is also considered. Each sentence (either labeled or unlabeled) changes to a vector of feature scores. Then each new coming sentence is observed in the feature space and gets two probabilities: how probable it is to be either labeled or unlabeled. These probabilities are calculated according to the position of new instance related to its labeled and unlabeled neighbors. We have applied the proposed model for improving training corpus of a SMT system. Also Farsi-English language pairs are selected as the base-line SMT system. We have sampled the best sentences that can improve the quality of our SMT system and send query for their translations. In this way the costly approach of making parallel corpus is alleviated. Finally, our experiments show significant improvements for sampling sentences by soft decision making in comparison to the random sentence selection strategy.**

Keywords-component; Active Learning, Statistical Machine Translation, Farsi and English pair Languages, Soft Decision Making, Kernel Based Distance, Density Based KNN.

I. INTRODUCTION

The predominant approach of Machine Translation (MT), Statistical approach (Brown et al. 1991), is based on parallel corpus. This means that a qualified Statistical MT (SMT) system needs a great amount of special input data as a bilingual parallel corpus. However, for most of the language pairs this is a kind of rare data, these languages being called scarce resource languages.

Producing parallel corpus is done either manually or automatically. In the first approach a noiseless corpus

will be prepared but this process of manually making a bilingual corpus by the help of human is very costly and time consuming. Automatically producing parallel corpus has no cost but the result is a noisy corpus. So, producing bilingual parallel corpora becomes a challenge in the process of developing a SMT system.

To reduce the cost of producing parallel corpus, the active learning (AL) approach proposes some methods that just queries for the translation of sentences which are essential for system improvement. These are samples that adding them to the training data improves the quality of system output.

The AL idea is one of the successful approaches for producing label for unlabeled instances in Machine Learning. This approach can be applied to different applications, e.g. Speech Recognition, Information Extraction, Classification and Filtering etc. For SMT story is a little different; as instead of labeling unlabeled data we are looking for the translation of source language sentences in the target language. The aim of AL for SMT is changing a monolingual corpus to a bilingual one.

Settles et al. (2009) have shown that in different NLP fields significant improvement can be gained during the process of annotating data by applying the AL idea. The strategies of how to send a query for new instances are divided to three categories see Fig. 1. (Settles et al., 2009);

1) Membership Making Synthesis,

2) Stream Based Sentence Sampling,

3) Pooled Based Sampling.

In the Membership Making Synthesis approach the system can send a query for any kinds of instances, even the system made instances. In this way, the queries can be so ambiguous that the process of label producing will be impossible, especially when the annotator is a human. Besides ambiguity, this approach is not proper for the tasks in which the order of tokens in the input data sequence is important, like NLP applications.

The second approach, Stream Based Sentence Sampling, is the process of visiting input data instances in sequential mode. It is proper for online data and situations that we have stream of data in which we have one data at time.

The third approach (Pooled Based Sampling) seems more qualified for NLP tasks. This approach investigates all instances at the same time and chooses

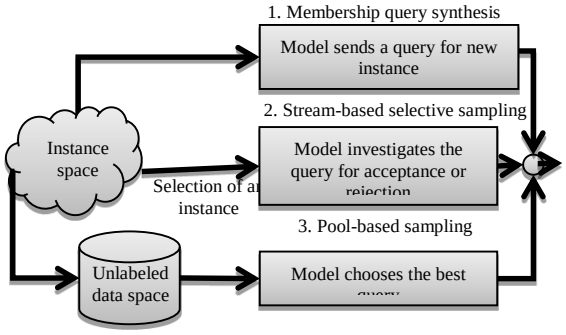

Fig. 1. Three main AL scenarios (Settles et al., 2009).

the best ones from the pool of instances, using a greedy search algorithm. It is especially proposed in the case of existence of a small set of labeled data in front of a large pool of unlabeled data.

In addition to the different strategies for sending queries, various strategies can also be chosen for scoring sentences. In most of them new instances are scored according to the amount of additional information that can add to the baseline system. For SMT the information of a sentence can be evaluated by the number of new phrases or new n-grams it can produce (Haffari et al., 2009a).

In this work, in addition to considering the amount of information of sentences, data distribution in the feature space is also considered. We have augmented the process of sentence scoring with the information about the density of input data space and the situation of unlabeled instances position in the space of labeled data. The results show that the performance of our algorithm has outperformed the random sampling process.

a. For better explaining the idea, process of our proposed algorithm is explained in separated steps:

b. Each sentence of both labeled and unlabeled instances are first mapped to the space of features vector,

c. Probability distribution of labeled and unlabeled instances are estimated based on labeled instances,

d. Each unlabeled instance is evaluated based on two distributions estimated in step (b), and space of unlabeled instances is sorted based on scores that are computed by using these distributions,

e. New instances from the sorted unlabeled data space are sampled,

f. Labels of selected samples are produced using human annotator,

g. The set of labeled instances is updated using new annotated instances.

The rest of the paper is structured as follows. In Section II, we describe related works in AL for NLP and particularly for SMT. In Section III, we introduce our new density-based AL algorithm in details. In Section IV, we analyze the results of our experiments. Finally, we conclude with a summary and an outline of further research in Section V.

II. RELATED WORKS

The use of AL ideas in the previous works can be categorized into two groups according to our project's requirements: applying these ideas to the NLP applications in general or using it for improving a SMT system in particular. Here we analyze these two groups separately. A common criterion among most of these researches is to compare the efficiency of AL algorithm with random sampling. Most of them have reported significant improvement compared to the random sampling approach (Tang et al., 2002; Haffari et al.,

2009a; Ambati et al., 2010; Chu et al., 2011; Li et al., 2012).

A. *Active learning in NLP*

The AL idea is applied to various applications in different fields of NLP (Settles, 2010), such as POS tagging (Engelson and Dagan, 1996; Ringger et al., 2007), parsing (Reichart and Rappoport, 2007), coreference resolution (Zhao and Ng, 2014), relation extraction (Qian et al., 2014), semantic annotation (Xu, 2014; Cui, 2014), word sense disambiguation (Chan and Ng, 2007; Zhu and Hovy, 2007), syntactical parsing (Hwa, 2004; Osborne and Baldridge, 2004), named entity recognition (Shen et al., 2004; Tomanek et al., 2007; Tomanek and Hahn, 2009) and sentiment analysis (Brew et al., 2010; Li et al., 2012; Xiao and Guo, 2013). Different works have used different techniques and also have attempted to measure the information of a new instance according to the different criteria. The main strategies which are mentioned in these works can be categorized as follow (Settles, 2010):

1 *Uncertainty Sampling* - Uncertainty sampling is The simplest and most common criterion for evaluating the instances that are first proposed for text classification applications (Lewis and Gale, 1994). The instances that we are more uncertain about their labels are probably the ones that are new according to the structure of the current system. Entropy is a good choice for measuring uncertainty in data. In addition Hwa (2004) and Settles et al. (2008) have proposed other measurements instead of the entropy criterion for more complicated data structures. They look for instances in which their best assigned labels have the least confidence among all the other unlabeled data. For SMT this criterion can be considered as the translation probability p(e|f), that is the baseline system probability provides for the translation of each unlabeled sentence f (Haffari et al., 2009a). Some other works use AL ideas in combination with semi-supervised learning approaches (Tomanek et al. 2009). Their strategy is also based on uncertainty. They apply AL for the sequence labeling task. In their approach, they ask human annotators to label only uncertain subsequences within the selected sentences while the remaining subsequences are labeled automatically based on the model trained on the available data produced during the previous AL iterations. Another work is (Vickrey et al. 2010), that uses seed words and iteratively expands this set by adding similar unlabeled words. In each iteration, AL suggests a series of candidate words which the user makes decisions to either accept it as a proper sample to produce its label or reject it. In this approach, they explore the space of similar words and send a query for the best candidate to be annotated by either a positive or negative label.

2 *Query-By-Committee* – This approach asks a group of baseline systems about the label of each instance. These sets of systems which are trained on labeled data are called committee. Committee members vote on the labeling of instances. The more disagreements between committee members about the proper label for each instance, the more informative that instance is (Seung et al., 1992). From SMT viewpoint, the implementation of this feature needs more than one translator.

3 *Expected Model Change* - Another strategy for choosing more informative sentences is looking for ones that adding them to the model causes to the greatest improvement in the model. For example, in discriminative probabilistic models that are based on regression the system improvement can be considered as the changes of the training gradient vector.

4 *Variance Reduction* - Cohn et al. (1996) propose some formula on how to decrease the future system errors by minimizing its variance. They have analyzed their idea for a regression based system. From the point of view of SMT, reduction of variance can be considered as enhancing current models' confidence by adding information about rare instances visited in the corpus. For that, queries that are more similar to the baseline model are proper for being queried. These instances will enhance the system models and reduce the system future errors.

5 *Estimated Error Reduction* - The previous criterion measures the system errors in relation to the system variance. In contrast, this approach attempts to directly estimate the reduction of system errors if the new unlabeled instance x is labeled and added to the baseline system. For different applications, various criteria are proposed (Settles et al., 2009). In

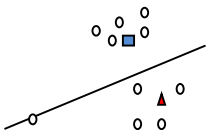

Fig. **2.** The impact of outliers on the quality of uncertainty sampling strategies.

SMT, the amount of BLEU scores (Papineni et al. 2002) improvement by adding each instance will be measured.

6 *Density-Weighted Methods* - The previous explained strategies are prone to the outlier instances. Fig. 2. explains the problem of outliers for binary classification while using the uncertainty sampling strategy. In Fig. 2. the instance located on the decision boundary of two classes is the most uncertain instance, proper to be added to the system but as it is an outlier no improvement in the quality of the class parameters will be identified. To get rid of these kinds of noises, the usage of some additional information which considers the input data distribution is proposed. Our model also considered this property by scoring sentences according to their situation on the space of features.

B. *Active learning in SMT*

Contrary to the various works that have applied AL to NLP tasks, the usage of AL for improving the SMT system is limited to a few works. The structure of a SMT system augmented with an AL module is described in Fig. 3. Probably the first work that has

suggested applying AL to MT is (Callison-Burch, 2003).

Their model, based on uncertainty strategy of AL, selects sentences that lead to uncertainty in MT systems. However, no experiments are reported. Some other works have attempt to optimize the size of training data such as (Eck et al., 2005) that have reduced the training set size by selecting instances that maximize n-gram feature coverage or (Lu et al., 2007) that have used TF-IDF based cosine score to select a subset of the parallel training sentences. Banerjee, et al. (2013) uses language models score for selecting informative sentences. They train two language models on well and badly translated sentences. The usefulness of a sentence is measured as the difference of its perplexities in these two language models. Biçici and Yuret (2014) have introduced a class of instance selection algorithms that use feature decay. In their model the training instances relevant to the test set are selected. As a new sentence selection strategy, Logacheva and Specia (2014) have proposed a new quality-based AL technique. The core idea of a quality-based AL technique is to select sentences that are likely to be translated incorrectly by the MT system. However, they have used a richer quality estimation metric which benefits from a wider range of features for estimating the correctness of automatic translation of a sentence.

The other quality-based AL technique for making an AL-SMT is proposed by Haffari et al. (2009a). In this work a wide range of efficient features are defined. Their suggested features for selecting informative sentences are based on improvement of a baseline system. They have shown how it is possible to improve a single SMT system by using the human translations of queries requested by AL. They have suggested some innovative features. These features are defined for evaluating sentences according to the amount of extra information which will be gained by adding each of them to the system. They have measured information of a sentence by counting the number of new phrases or n-grams it can produce in comparison to the phrases or n-grams observed in the baseline corpus. Some other features which consider the translation quality are also proposed. On the basis of the previous work, Du, et al., (2014) introduced a length penalty factor into the phrase-based sentence selection strategy to penalize the short sentences. The penalty factor is updated in each iteration.

Haffari et al. (2009b) have improved the multilingual SMT systems by extracting the instances which are more likely to improve the translation quality. They have proposed a notion for better handling the usage of the AL idea to improve a small bilingual corpus with big monolingual data. They have suggested two approaches: self-training and co-training. In self-training approach, the corpus is enlarged with the human translated sentences plus the system's noisy translated output while in co-training each of the SMT systems is improved by the human translated instances, in addition to the other systems' translation output.

The other work reported in the AL-SMT subject is (Ambati et al., 2010), which proposes a new approach for enabling automatic translation for languages with low resources. They use the Active Crowd Translation (ACT) idea that is a combination of AL and Crowd-sourcing ideas. In this work, they choose more informative sentences while trying to reduce the cost of translation through Amazon Turk. Their work has two parts: the first part is based on AL for sentence sampling and the second part is based on crowd-sourcing for finding good translation among Turker's suggestions. In AL, the strategy of choosing new instances is based on the produced phrases by each sentence. Thus, the sentences which produce most representative n-grams that have not been seen yet in the bilingual corpus are chosen. Finally, each sentence is scored according to two factors: density (frequency of phrases in the labeled data) and uncertainty (the number of new phrases that a sentence can produce).

Bloodgood et al. (2010) have a new point of view to the problem of applying AL for improving a SMT system. All previous works which have used AL in the field of SMT are trying to solve the problem of scarce resources. Thus, they start from an extremely small set of seed data and in each iteration, add a very tiny amount of data during the AL process. However, Bloodgood et al. (2010) have demonstrated how to apply AL in situations where a large corpus is available. Their goal is to buck the trend of diminishing return. The diminishing circumstance always occurs from some iteration in the AL process.

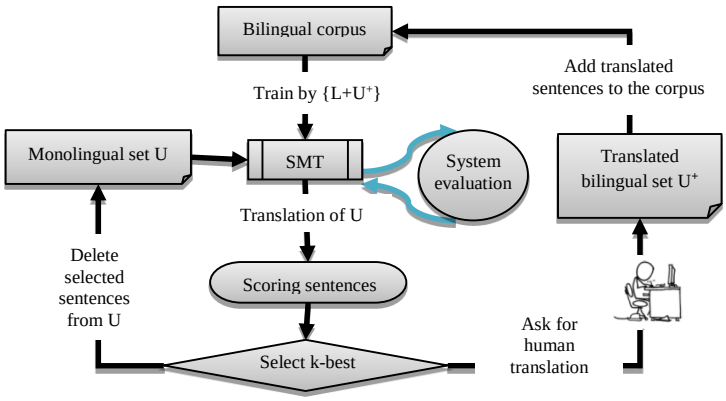

Fig.**3** The structure of a SMT system augmented with an AL module

They tend to make highest-performing MT systems while keeping annotation costs low. Actually, the algorithm of Bloodgood et al. (2010) starts when the other approaches stop their investigations. After sampling, they gather annotations via the Amazon Mechanical Turk.

U = Monolingual corpus,
L = Bilingual corpus,
For t=1,2,…
  For all $S \in U$
    U^s = Score all S in U by $\{f_1^n\}$
    L^s = Score all S in L by $\{f_1^n\}$
    featureSpace = $\{U^s \bigcup L^s\}$
    N = {k nearest neighbor of S in featureSpace }
    $\delta = 2$

$$P_u(s) \propto \sum_{n \in \{N \cap U\}} \frac{\exp\left(-1 \times (d(s,n))^2\right)}{\delta}$$

$$P_l(s) \propto \sum_{n \in \{N \cap L\}} \frac{\exp\left(-1 \times (d(s,n))^2\right)}{\delta}$$

    If $(P_u(s) > P_l(s))$
      Consider s as probable unlabeled instance,
    else
      Consider S as probable labeled instance and select the k-best from probable unlabeled instances and add them to L
  end for
end for

Fig. **4.** The pseudo code of density- based KNN for sampling sentences.

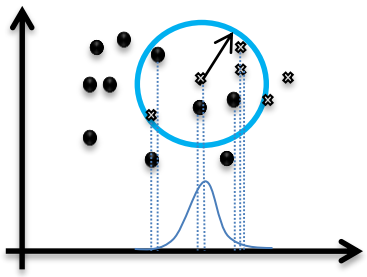

Fig. **5.** KNN density based algorithm which uses a kernel on top of each candidate data.

There are also some other works that use AL idea for improving an SMT quality but not by expanding the training corpus. Dara, et al. (2014) apply AL for capturing human post-editing outputs as early as possible to incrementally update SMT models to avoid repeat mistakes. González-Rubio and Casacuberta (2014) propose a cost-sensitive AL framework for computer-assisted translation. They optimize the number of human supervision and difficulty of his/her attempt. Thus, they focus the user effort to those translations which user supervision considered as more "informative".

In our previous work (Bakhshaei et al., 2010), we have studied on how to apply AL idea for expanding Farsi-English corpus. In this work the efficiency of some of the proposed features were investigated on Farsi-English pair of languages. The results showed the expanded corpus in this way can improve the quality of an SMT system.

Our work completely differs from previous works. Features that are used in this work are similar to Haffari et al. (2009a), but we have proposed a new approach for applying these features. Our method is completely different; we map a sentence to a vector of features and observe instances in the feature space. The contribution of our work is that we sample sentences by considering density of data in the features space. In this way the strategy of AL in our work is more similar to the Density-Weighted strategy. Sampling in our model is done through a kind of soft decision making process.

This paper contains new idea for AL sampling. More details are presented below:

- New instance description- In this paper in contrast to (Bakhshaei et al, 2010), we use groups of features for describing a sample in the instances space, while in the similar previous works each of the features is separately used for scoring unlabeled instances or a in some others, mixture of features is used.
- New sampling approach- Also in this paper we propose a new approach for sampling the unlabeled instances. In this approach we consider the location of instances in the feature space and through a soft decision making. We count how much it is probable for an instance to be labeled or unlabeled. We prefer instances which are located near to the dense location of the labeled instances space and are far from the unlabeled instances. The classification of instances is based on a modified K-nearest neighbor algorithm. The details are explained in the next section.

III. KNN DENSITY BASED ACTIVE LEARNING

It is shown that applying AL approaches reduces the cost of annotating unlabeled samples (Settles et al., 2008; Ambati et al., 2010), also for application like MT, applying AL strategies requires a smaller number of sentences to reach a desired performance thereby reducing cost of acquiring data (Ambati et al., 2010). In this way, training a system with less data is possible while accuracy does not decrease. Thus, in the case of scarce data resources we do not have to gather a great amount of bilingual data; just a customized size of informative sentences is enough for training a qualified system. Finding proper choices are possible by defining appropriate features for evaluating the information of each sentence. A comprehensive group of features is defined by (Haffari et al. 2009a). Their suggested features are applied for sampling in two modes. In the first mode, they have considered the score of each individual feature for ranking monolingual corpus sentences and then choose instances according to the value of each separated feature heuristically:

$$\forall s \in U : score(s) = score_{f_i}(s) \qquad (1)$$

In equation (1), s stands for any unlabeled instances from U set. $score(.)$ is score of a sentence and $score_{f_i}(.)$ is a function that evaluates each sentence according to value of feature f_i. In the second mode they combine the score of all the features and use it as a single mixture feature.

For this aim they have used the weighted combinations of all the features (WCF) score in addition to a new approach which is called Hierarchical Adaptive Sampling (HAS) for translation. In WCF the effect of each feature is controlled by assigning a weight to it. The weight vector must be tuned on the corpus. Then the score of each sentence is counted according to a mixture of features:

$$\forall s \in U : score(s) = \sum_{f_i} \alpha_i \times score_{f_i}(s) \quad (2)$$

In equation (2), parameter α_i is the weight of feature f_i.

The new HAS algorithm, samples sentences of U (unlabeled data) while building a hierarchical cluster on the ranked sentences.The strategies explained above for sentence sampling are all a kind of hard decision making, while an attempt has been made to make a soft decision making, in this paper. Our soft decision making is based on assigning each instance two probabilities of how much it is probable to be either a labeled or an unlabeled data. In this definition a sentence that is a proper choice for being sampled is the one that is more unlabeled than labeled instance. Also, we have forced the system to consider the distribution of the unlabeled instances in the input space by allowing for density of the data in the scoring process in addition to measuring the amount of their information. We have used the nonparametric KNN algorithm for classifying sentences while using the proposed features in previous works for featuring sentences. Each sentence either labeled (L) or unlabeled (U) is changed to a vector of values related to the score of these features:

$$\forall s \in \{L \bigcup U\}: \quad s \rightarrow < score_{f_1}(s), ..., score_{f_n}(s) > \quad (3)$$

In equation (3), s stands for any instance either labeled or unlabeled, $score_{f_i}$ is the same as what is defined in equation (1) and $< score_{f_1}(s), ..., score_{f_n}(s) >$ is a vector that its value in index i is equivalent to the score of feature i.

The pseudo code of our algorithm can be seen in Fig. 4. In this algorithm we look at data in the feature space containing both labeled and unlabeled data. Then proper sentences are selected to ask for their translations. As explained above our soft decision making is carried out by assigning both U-probability ($P_u(.)$) and L-probability ($P_l(.)$) to each instance s according to the equations 1 & 2. In this way, we count how probable it is for a sentence s to be either a labeled or unlabeled instance.

$$P_u(s) \propto \sum_{n \in \{N \cap U\}} \frac{\exp\left(-1 \times (d(s,n))^2\right)}{\delta} \quad (4)$$

$$P_l(s) \propto \sum_{n \in \{N \cap L\}} \frac{\exp\left(-1 \times (d(s,n))^2\right)}{\delta} \quad (5)$$

In equations (4) and (5) N is the set of nearest neighbors to instance s. $N \cap L$ and $N \cap L$ are the sets of labeled and unlabeled instances in the neighbor of s respectively. Also, $d(.,.)$ is a distance metric and we have considered it as a Euclidean distance. In this equation, an effect zone is allotted to each sentence by using a kernel $k(x) = \exp\left(-\text{distance}(x, x')^2\right) / \delta$ on top of that candidate. Thus, the closer neighbors are considered more important than the neighbors in the longer distances. The effective domain of this zone is controlled by changing δ parameter in formula (4) and (5). This kernel distance metric forces the algorithm to choose candidates from the regions where more unlabeled data are allocated and this is equivalent to consider the density of unlabeled data space in scoring the sentences process. The graphical view of the algorithm is shown in Fig. 5.

IV. EXPERIMENTS

To show the qualification of our proposed algorithm we set up different tests. For this means we used a bilingual Farsi-English corpus which is a part of the corpus produced by (blind-a). The corpus is separated to two sets, one of which is used as a bilingual corpus (labeled data) and the other as monolingual corpus (unlabeled data). The standard phrase-based model that we used for training is the Moses system (Koehn et al., 2007) in which we used

Table 1. Data statistic for labeled data (or Training set) (a), unlabeled data (b) and test and Dev (Development) sets (c).

	Labeled Data			
	Sentence	**Running words**	**Singleton**	**Lexicon**
English	**5000**	**49373**	**475**	**1314**
Farsi	**5000**	**46435**	**1005**	**2318**

(a)

	unlabeled Data			
	Sentence	**Running words**	**Singleton**	**Lexicon**
Farsi	**18145**	**170045**	**2248**	**4742**

(b)

	Sentence	**Running words**	**Singleton**	**Lexicon**
Dev set	**2001**	**15127**	**1987**	**3277**
Test set	**3500**	**28597**	**2532**	**4551**

(c)

default values for all of the parameters. All experiments use a 4-gram language model trained on the Farsi side of our training corpus using SRILM (Stolcke, 2002) with Kneser-Ney smoothing (Kneser, and Ney, 1995). To tune feature weights in minimum error rate (Och, 2003) training, we used a development set of 2,001 sentence pairs, and we evaluate performance on a test set of 3,500 single-references. For more information of the data see Table 1.

A baseline SMT system is trained on the bilingual corpus (L) and the density-based KNN algorithm is applied to the monolingual corpus (U) for sentence sampling. Note that U is just the source side of the main corpus (the target side is ignored), see Table 1-b. The sentences are featured according to the defined features: Geom-n-gram, Arith-n-gram, Geom-Phrase, Arith-Phrase.

These features have no problem with unlabeled data but evaluating labeled data with these features is slightly a vague process since each of these features compares some characteristics of the current sentence with the quality of it in the labeled data set. For example, the Geom-n-gram feature is the geometric average of relative frequency of n-grams of the current sentence in comparison to the visited n-grams of L:

$$\phi_g^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \qquad (6)$$

In equation (6), X_s^p is the set of all n-grams which can be extracted from sentence s and $P(x|U,n)$ is the probability of n-gram x in U. Parameter w_n is the weight of each n-gram. In this work we have considered equivalent weight for all n-grams.

Evaluating unlabeled sentences by this feature (equation (6)) has no problem but evaluating the sentences of L faces to some problem; the feature will compare L sentences with themselves, thus, this equation is always 0 for each L sentence. Haffari et al. (2009a) have solved this problem by applying leaving-one-out algorithm for featuring L sentences.

As leaving-one-out algorithm is a very time consuming process, we have used n-fold algorithm instead. Used features are explained briefly in this section.

A. Phrase features

In this feature we consider the amount of new phrases a sentence can add to the system. Thus, we count the relative frequency of the phrases produced by this sentence to the total phrases seen in the current corpus.

Finally, the score of the sentence can be measured by averaging all of the sentence phrases. We can use either geometrical or arithmetical means. Each of these options is considered as a separate feature in equations (7) & (8).

$$\phi_g^p(s) := \left[\prod_{x \in X_s^p} \frac{P(x|U)}{P(x|L)}\right]^{\frac{1}{|X_s^p|}} \qquad (7)$$

$$\phi_a^p(s) := \frac{1}{|X_s^p|} \sum_{x \in X_s^p} \frac{P(x|U)}{P(x|L)} \qquad (8)$$

In the formulas (7) and (8) X_s^p is the set of all possible phrases that can be extracted from the current sentence s. The probability of each sentence in set D∈{U, L}, $P(x|D)$, is computed according to $P(x|D) = \frac{count(x)+\varepsilon}{\sum_{x \in X_s^p} count(x)+\varepsilon}$, ε is the smoothing factor and is set to a small value.

B. Ngram features

The second feature is the amount of n-grams that each sentence can produce. The average frequency of produced n-grams of the current sentence relative to the n-grams observed in the base-line corpus is defined as the Ngram feature. Just like the previous feature, arithmetic or geometrical mean of scores are considered as two separated features in equations (9) & (10).

$$\phi_g^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \qquad (9)$$

$$\phi_a^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \frac{P(x|U,n)}{P(x|L,n)} \qquad (10)$$

In equations (9) and (10), X_s^n is the set of all n-grams that can be produced from sentence s Probability function $P(x|D,n)$ that D∈{U, L} is the probability of x in the n-grams extracted from D set. w_n is the weight of each n-gram that is considered the same as equation (6).

C. Reverse model feature

In this feature we count how accurate each sentence is translated in the SMT system trained on current L set. It is measured by passing sentences from the target-to-source system and re-passing the result from the source-to-target system.

The similarity of the final result with the main sentence, which is measured by BLEU score, reveals the confidence of the system.

D. Translation confidence feature

The negative of translation probability of a sentence, $-p(f|e)$, is considered as the confidence measure of current system about its translation. Actually, the translation probability shows how good a sentence is translated in a system and reveals if the system has enough knowledge for translating the sentence.

So, we choose the sentences that have achieved the worse translation probability. These are the ones that the system is in problem for translating them and needs to learn more about them.

V. EXPERIMENTAL RESULTS

The results of applying the features (explained in section IV) separately for choosing sentences are shown

in Fig. 6. Two probabilities are assigned to each sentence s in U:

1. How much it is likely that *s* to be unlabeled and

2. How much it is likely that *s* to be labeled.

Then in an iterative algorithm, a list of 5000-best sentences is chosen among the most unlabeled data. The chosen sentences are paired with their translations produced by human translator and are added to the baseline corpus L.

Finally, the SMT system is retrained on this new parallel corpus and the qualification of the system is estimated by translating the test set (see Table 1-c) and counting the BLEU score of the results.

The results of applying our algorithm to the data show significant improvement. BLEU scores are reported in Table 2 and for simplifying the comparison of the results, Fig. 7. has depicted the BLEU scores in each iteration.

In Fig.6. unlabeled samples are scored according to formula (1) and best score sentences with their translations are added to the parallel corpus.

In each iteration i, 5000 sentences are selected and added to the parallel corpus. Almost each step improves the BLEU score of the results but some fluctuations has occurred.

We have repeated the same circumstances in Table 2. But instead of evaluating unlabeled samples using formula (2), we used Formula (3).

First each sentence is changed to a vector in the space of features, then through a soft decision making we judge if the sample is probable to be labeled or unlabeled using formula (4) and (5).

Finally, algorithm of Fig. (4) is applied to select the best choices to be added to the training corpus.

By observing the BLEU scores in each iteration, it is seen that the density based AL has led to more confident results and the fluctuations in the BLEU score which is common in the most of the results shown in Fig. 6., has not occurred.

According to the diagram of Fig. 7. that depicts the BLEU score in each iteration for sampling with the help of the proposed algorithm outperforms the random sentence sampling process.

VI. CONCLUSION AND PERSPECTIVES

In this paper we had a short review on the AL idea by considering its application in different NLP tasks. We discussed AL details on both sending query and scoring sample processes.

In particular, we have discussed AL for SMT by going over the reported research in this realm. We have proposed a new version of AL for sampling new sentences for SMT applications.

This algorithm defines a neighborhood around each candidate sentence and assigns the two probabilities of being a labeled or unlabeled data.

This soft decision making also considers the density of unlabeled data in the input space by putting a kernel on top of each candidate sentence. The results show that the algorithm can work better than random sampling.

The proposed algorithm can be improved in some parts that we intend to set as the plan of our future works.

The number of used features and their quality of features must be studied deeper. Quality of a feature is how successful a feature is to select the best samples.

Also the amount of information which is hold by chosen samples must be considered. On the other hand the number of features can be optimized for improving the time complexity.

Feature selection approaches might be helpful for improving the proposed model. The efficiency of the distance metric that is based on Euclidean must be considered. Also, the qualification of this metric for using as the similarity criterion of the sentences in the feature space must be investigated.

The proposed model is capable of more extended uses. For example it is proper to be used in the other applications such as phrase table or corpus filtering for removing noises or optimizing size without changing performance also the model is useful for domain adaptations.

However for making model to be applicable for these new aims, features must be customized.

Table 2. The BLEU score of the SMT systems that are trained on expanded corpus using the Density-based AL model and random sampling approach.

Iteration #	Expanded Corpus size	Density-based AL model BLEU[%]	Random sampling model BLEU[%]
1	5000	0.1486	0.1419
2	9735	0.1624	0.1418
3	14298	0.1608	0.1601
4	18519	0.166	0.1578

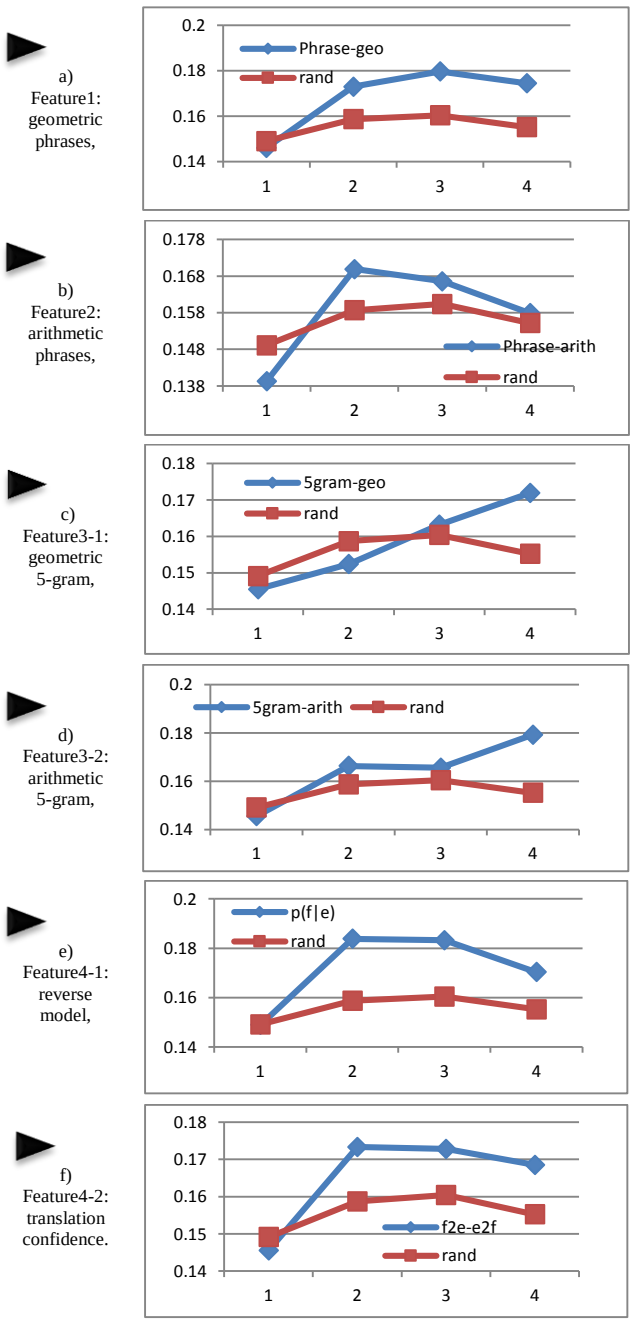

Fig. **6.** The BLEU score of the SMT systems trained on expanded corpora by sampled sentenced by the help of the features defined in section 4. "rand" in the above figures is the result of random sentence selection approach.

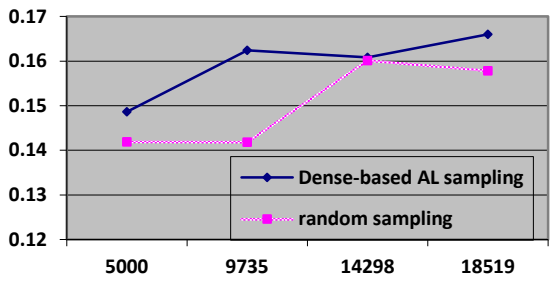

Fig. **7.** Comparison of the BLEU score of our model and the random sampling model.

REFERENCES

Ambati, V. Vogel, S. and Carbonell, J., 2010. Active learning and crowd-sourcing for machine translation. Language Resources and Evaluation (LREC).

Bakhshaei, S., Khadivi, S. and Riahi, N. , 2010. Farsi-German statistical machine translation through bridge language. Telecommunications (IST5th) International Symposium on. IEEE.

Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J., 2013. Quality estimation-guided data selection for domain adaptation of smt. MT Summit XIV: proceedings of the fourteenth Machine Translation Summit, 101-108.

Biçici, E., and Yuret, D., 2014. Optimizing instance selection for statistical machine translation with feature decay algorithms. IEEE. ACM Transactions On Audio, Speech, and Language Processing (TASLP).

Bloodgood, M., Callison-Burch, C., 2010. Bucking the trend: large-scale cost-focused active learning for statistical machine translation. ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics Pages 854-864.

Brew, A., Greene, D., and Cunningham, P., 2010. Using crowdsourcing and active learning to track sentiment in online media. ECAI'2010: 145–150.

Brown, Peter F., Stephen A. Pietra, D., Vincent J. Pietra, D., Robert L. Mercer, 1991. Word-sense disambiguation using statistical methods. In Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. pp. 264-270.

Callison-Burch, C., 2003. Active Learning for Statistical Machine Translation. PhD Proposal, Edinburgh University.

Chan, Y. S., and Ng, H. T., 2007. Domain adaptation with active learning for word sense disambiguation. ACL'2007. 590.

Chu, W., Zinkevich, M., Li, L., Thomas, A., and Tseng, B. 2011. Unbiased online active learning in data streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 195-203). ACM.

Cohn, D. Ghahramani, Z. and Jordan. M. I., 1996. Active learning with statistical models. Journal of Artificial Intelligence Research, pages 129–145.

Cui, S., Dumitru, C. O., and Datcu, M., 2014. Semantic annotation in earth observation based on active learning. International Journal of Image and Data Fusion, 5(2), 152-174.

Dara, A., Genabith, J., Liu, Q., Judge, J., Toral, A., 2014. Active Learning for Post-Editing Based Incrementally Retrained MT. EACL 2014, 185.

Du, J., Wang, M., and Zhang, M., 2014. Sentence-Length Informed Method for Active Learning Based Resource-Poor Statistical Machine Translation. In Natural Language Processing and Chinese Computing (pp. 91-102). Springer Berlin Heidelberg.

Eck, M., Vogel, S., and Waibel, A., 2005. Low cost Portability for statistical machine translation based on n-gram frequency and TF-IDF. In IWSLT (pp. 61-67).

Engelson and I. Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. ACL'1996: 319–326.

González-Rubio, J., and Casacuberta, F., 2014. Cost-sensitive active learning for computer-assisted translation. Pattern Recognition Letters, 37, 124-134.

Haffari, G., and Sarkar, A., 2009b. Active Learning for Multilingual Statistical Machine Translation, Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 181–189.

Haffari, G., Roy, M. and Sarkar, A., 2009a. Active learning for statistical phrase-based machine translation, In NAACL.

Hwa, R., 2004. Sample selection for statistical parsing. Computational Linguistics, 30(3): 253–276.

Jabbari, F., Bakhshaei, S., and Ziabary, S. M. M., Khadivi, S. 2012. Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus. In The Fourth Workshop on Computational Approaches to Arabic Script-based Languages (p. 17).

Kneser, R., and Ney, H. 1995. Improved backing-off for m-gram language modeling. In Acoustics, Speech, and Signal Processing. ICASSP-95., 1995 International Conference on (Vol. 1, pp. 181-184). IEEE, 1995.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177-180). Association for Computational Linguistics.

Lewis, D. and Gale, W., 1994. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12.

Li, S., Ju, S., Zhou, G., and Li, X. 2012. Active learning for imbalanced sentiment classification. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 139-148). Association for Computational Linguistics.

Logacheva, V., and Specia, L., 2014. A quality-based active sample selection strategy for statistical machine translation. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

Lü, Y., Huang, J., and Liu, Q., 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. InEMNLP-CoNLL (pp. 343-350).

Och F. J., 2003. Minimum error rate training in statistical machine translation. Association for Computational Linguistics. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Osborne, M., and Baldridge, J., 2004. Ensemble based active learning for parse selection. HLT-NAACL' 2004: 89–96.

Papineni, K. Roukos, S. ToddWard, and Zhu, W. J., 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Qian, L., Hui, H., Hu, Y., Zhou, G., Zhu, Q., 2014. Bilingual Active Learning for Relation Classification via Pseudo Parallel Corpora. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 582–592, Baltimore, Maryland, USA, June 23-25 2014.

Reichart, R., and Rappoport, A., 2007. An ensemble method for selection of high quality parses. In ACL (Vol. 7, pp. 408-415).

Ringger, E., McClanahan, P., Haertel, R., Busby, G., Carmen, M., Carroll, J., Seppi, K. and Lonsdale, D., 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In Proceedings of the Linguistic Annotation Workshop (pp. 101-108). Association for Computational Linguistics.

Settles, B., Craven, M., and Friedland, L. 2008. Active learning with real annotation costs. In Proceedings of the NIPS workshop on cost-sensitive learning (pp. 1-10).

Settles, B. and Craven, M., 2009. An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1069–1078.

Settles, B., 2010. Active learning literature survey. University of Wisconsin, Madison, 52(55-66), 11.

Seung, H. S. Opper, M. and Sompolinsky, H., 1992. Query by committee. In Proceedings of the ACMWorkshop on Computational Learning Theory, pages 287–294.

Stolcke A., 2002. SRILM -- An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver.

Tang, M., Luo, X., & Roukos, S. 2002. Active learning for statistical natural language parsing. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 120-127). Association for Computational Linguistics.

Tomanek, K., and Hahn, U., 2009. Semi-supervised active learning for sequence labeling. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1039-1047). Association for Computational Linguistics.

Vickrey, D., Kipersztok, O., and Koller, D., 2010. An active learning approach to finding related terms. In Proceedings of the ACL 2010 Conference Short Papers (pp. 371-376). Association for Computational Linguistics.

Xiao, M., and Guo, Y., 2013. Online Active Learning for Cost Sensitive Domain Adaptation. CoNLL-2013, 1.

Xu, K., Liao, S. S., Lau, R. Y., and Zhao, J. L., 2014. Effective Active Learning Strategies for the Use of Large-Margin Classifiers in Semantic Annotation: An Optimal Parameter Discovery Perspective. INFORMS Journal on Computing, 26(3), 461-483.

Zhao, S., and Ng, H. T., 2014. Domain Adaptation with Active Learning for Coreference Resolution. In Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL (pp. 21-29).

Zhu, J. B., and Hovy, E., 2007. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. EMNLPCoNLL'2007: 783-790.

Somayeh Bakhshaei received her B.Sc. degree in computer science from Sharif University of Technology (2005) and her M.Sc. degree in Computer Engineering (Artificial Inteligence) from Alzahra University (2010). She is currently a Ph.D. student in Computer Engineering at the Amirkabir University of Technology.Her research interests include Natural Language Processing, Statistical Machine Translation and Information Retrieval.

Reza Safabakhsh received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1976 and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Tennessee, Knoxville, in 1980 and 1986, respectively. He worked at the Center of Excellence in Information Systems, Nashville, TN, USA, from 1986 to 1988. Since 1988, he has been with the Computer Engineering Department, Amirkabir University of Technology,Tehran, Iran, where he is currently a Professor and the director of the Computer Vision Laboratory. His current research interests include machine learning, natural language processing, and computer vision. Dr.Safabakhsh is a member of the IEEE and several honor societies, including Phi Kappa Phi and Eta Kapa Nu. He was the founder and a member of the Board of Executives of the Computer Society of Iran, and was the President of this society for the first 4 years.

Shahram Khadivi received the B.S. and M.S. degrees in computer engineering from Amirkabir University of Technology, Tehran, Iran, in 1996 and 1999, respectively, and the Ph.D. degree in computer science from RWTH Aachen University, Aachen, Germany, in 2008. From September 2008 to January 2015, he was assistant professor at Computer Engineering department at Amirkabir University of Technology. His research interests include statistical machine translation, computational natural language processing, information retrieval, machine learning, and data analysis.