Artif Intell Rev (2016) 46:145–166
DOI 10.1007/s10462-016-9460-3

Learning sign language machine translation based on elastic net regularization and latent semantic analysis

Mehrez Boulares[1] · **Mohamed Jemni**[1]

Published online: 14 January 2016

Abstract In this paper, we present a new sign language machine translation approach based on regression method. The aim of this work is to improve the translation quality and accuracy of existing regularized regression methods. Our approach represents a methodological foundation for small-scale corpus domains such as the Sign Language Machine Translation field. Our method is based on the Elastic net regularization using linear combination of the L1 and L2 penalties of the lasso and ridge methods. We show that using both the de-bruijn graph with the Latent Semantic Analysis technique in the decoding process improves the translation results. The system is experimented on American Sign Language parallel corpora containing 300 sentences and assessed by BLEU, METEOR, NIST and F1-MESURE machine translation evaluation metrics. We obtained good experimental results compared to classical phrase based approach i.e MOSES framework. Also our approach improved the translation results compared to LASSO and RIDGE regression approaches.

Keywords Machine translation · Sign language · Elastic net regularization · Latent semantic analysis

1 Introduction

Machine learning technology has progressively extended its application domain to reach Machine Translation (MT) field. It has been improved by the emergence of effective statistical methods such as phrase-based MT (statistical MT). However, phrase-based MT is considered as a powerful statistical methods based on automatic train systems from a very large translated text sources (called parallel corpora). It relies on heuristics for an unsupervised high quality

✉ Mehrez Boulares
mehrez.boulares@gmail.com

Mohamed Jemni
mohamed_jemni2000@yahoo.fr

[1] Research Laboratory of Technologies of Information and Communication & Electrical Engineering (LaTICE), University of Tunis, Tunis, Tunisia

word alignments based on IBM models (IBM Model 1, 2, 3, 4) and word frequency. The effectiveness of this system depends on the quality of the alignment and on the size of the parallel training corpora. In other words, the performance of statistical based MT is closely related to the availability of very large parallel corpora. This is not the case for sign language data.

Although the quality and the availability of sign language (SL) corpora have been improved greatly in the past few years Neidle and Sclaroff (2002), Efthimiou (2007). The majority of existing sign language corpora are focused on video annotation such as the NCSLGR Corpus (National Center for Sign Language and Gesture Resources) and the BSL (British Sign Language Project). In the literature, there are some attempts to create parallel textual SL corpora such as the German RWTH-PHOENIX-Weather corpus. But, in general, there is a lack of multilingual large parallel corpora for Sign Languages. This represents a significant obstacle for sign language researches Efthimiou et al. (2009) in particularly for Sign Language MT.

This work is part of WebSign project Boulares and Jemni (2012) that aims to translate text to sign language animation using 3D virtual person. WebSign project is composed mainly from Machine Translation module and animation module. In this paper, we describe the machine translation module that aims to translate a manually pre-treated transcription form of English text to American Sign Language (ASL) textual transcription including Signing Space information (presented in Sect. 3). The choice of the ASL transcription is related to the availability in the literature of relevant works studying the structure of ASL sentences Neidle et al. (2000), Stokoe (1978) which is, in the general, not the case for many other languages. In this context, we created an ASL corpus composed of 300 parallel phrases to train our MT approach as well as 100 different parallel phrases to test the translation process.

However, in the MT field, phrase-based and regression-based MT are considered as the most relevant techniques. The phrase-based MT is well known as a powerful statistical method requiring very big training data to give good translation results. The translation process of the regression based MT relies on linear regression learning of word-to-word feature mapping. For new unknown input phrase (feature vectors), the learned linear regression model is used to predict the target feature vector. Once the target feature vector is obtained, a multi-graph search is used to find all the possible target words whose mappings correspond to the translated feature vector. This step is called decoding process. In order to argue our MT choice, we conducted experimentation based on these two techniques. We proved experimentally that our regression based MT performs better than phrase based MT in the context of small-scale corpora.

The main contribution of this work involves three aspects. The first one aims to use the existing regression approach "Elastic Network" Hastie et al. (2006) which is known as an improvement of the l1-norm Lasso and L2-norm Ridge in term of fitting accuracy R squared score of the regression function Colin et al. (1997). The second one focuses on the application of the Latent Semantic Analysis (LSA) in the decoding process after learning the regression function by Elastic Net method. The third aspect is related to the application of this approach to small-scale ASL parallel corpora with simple ASL phrase structure (Subject Object Verb form).

In this context, we created a small-size parallel corpora composed of 300 manually pre-treated English sentences in order to obtain suitable representation of ASL. We used the N-Spectrum Weighted word Kernel Leslie and eskin (2002), Watkins (2000) to generate feature vectors mapping of both source and target 2-grams. To learn the function that maps source to 2-grams target, we used and compared the l1-norm (LASSO) and the L2-norm (RIDGE) to the Elastic Net method in order to maximize the R squared and therefore to

improve the translation accuracy. As a solution to the pre-image problem (decoding process), we used the De-Bruijn Multi-Graph search applied on the 2-grams target. In order to improve the classical decoding process that uses Language Model, we used LSA searching method. We conducted set of experimentation to compare our approach with others i.e MT framework MOSES, LASSO and RIDGE based regression MT.

The remainder of this paper is organized as follows. In Sect. 2, we present the related works. Section 3 is devoted to describe the sign language data. Section 4 is dedicated to present our approach. Section 5 presents our experimentation and main results we obtained. Finally, the conclusion and some perspectives.

2 Related work

One of the main problems of machine translation is how to find the most likely translation of a source sentence from a set of training sentences. The goal is to find significant relationships between source and target language. However, due to the complexity of the problem, it is not easy to express these relationships as a set of rules. In other words, there are no general rules that can generate high quality translation for new unknown input sentences.

In fact, the majority of relevant research works are based on statistical MT and regression based MT. The work of Koehn et al. (2003), Koehn and Hoang (2007) was focused on statistical models whose parameters are derived from the automatic analysis of a set of bilingual phrases. This work relies on the search of the highest probability translation within a number of choices in order to find the most likely translation of an input text. This technique gives good results in term of translation quality based on very big bilingual text corpora. The quality of the results is closely related to the size of the training parallel corpora and to the quality of the unsupervised alignment. However, the translation cannot be of high quality if this technique uses a reduced training set such as on sign language corpora.

The regression based MT relies on linear regression which is classified as one of the methods of multivariate analysis that deal with quantitative data. The main objective of this method is to seek a linear mapping between one or more quantitative source variable and one or more quantitative target variables. However, MT method deals with the problem of mapping sentences x from a source language $X^*$to a target language Y^*. Formally, let X and Y correspond to the token sets used to represent source and target N-Gram, then a training sample of m input N-Grams can be represented as: $(X_1,Y_1)\ldots(X_m, Y_m) \in X^* \times Y^*$, where (x_i, y_i) corresponds to a pair of source and target language token string. Input N-Gram in $X^*$are mapped via φ_x to feature space F_x and the output string are mapped to F_y via the mapping φ_y. The mapping can be defined implicitly by a positive symmetric Kernel K_x and K_y associated with the mappings φ_x and φ_y. Our goal is to find a mapping f: $X^* \rightarrow Y^*$ that can convert a given set of source phrases to a set of target phrases that share the same meaning in the target language. Our objective is to predict $F_y$with target features $(K > 1)$, based on a multiple regression problems which could be done by introducing a different set of basis functions for each feature. In other words, a multiple regression technique can be used to

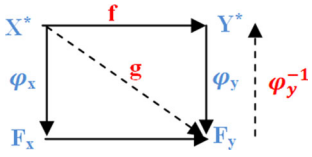


Fig. 1 The String-to-string mapping

learn and to estimate the mapping g from X^* to $F_y$based on its pre-image set ${\varphi^{-1}}_y$. Figure 1 depicts the scheme of the translation process presented in Cortes work Cortes et al. (2007).

However, the common approach used is to use the same set of basis functions in order to model all the target features. functions in order to model all the target features.

$$\varphi(y) = W\varphi(x)$$

$$\begin{pmatrix} | & | \\ | & | \\ \varphi_{y(1)} \ldots \varphi_{y(n)} \\ | & | \\ | & | \end{pmatrix} = \begin{pmatrix} - & w_1 & - \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ - & w_q & - \end{pmatrix} \begin{pmatrix} | & | \\ | & | \\ \varphi_{x(1)} \cdots \varphi_{x(n)} \\ | & | \\ | & | \end{pmatrix} + \begin{pmatrix} | & | \\ | & | \\ \varepsilon_1 \ldots\ldots \varepsilon_n \\ | & | \\ | & | \end{pmatrix} \quad (1)$$

This problem can be solved based on the minimization of the sum of squared differences (SSD) in $\varphi(y)$ on S where S is a set of bilingual sentence pairs. $S = \{(x_i, y_i) : w_i \epsilon X^*, y_i \epsilon Y^*, (i = 1...m)\}$. This solution [Eq. (2)] is known as Ordinary least Squares MCO or L^2 Norm that aims to learn the linear operator W in Eq. (1):

$$min||WM_{\varphi(X)} - M_{\varphi(Y)})||_F^2 \quad (2)$$

where $M_{\varphi(X)} = [\varphi(x_1), ..., \varphi(x_n)]$, $M_{\varphi(Y)} = [\varphi(y_1), ..., \varphi(y_n)]$ with M is a matrix and $||.||_F$ denotes the square root of the sum of the absolute squares of the elements known as the Frobenius norm (matrix norm). The minimal least squares estimator is given by:

$$W = (M_{\varphi(x)}^T \ M_{\varphi(x)})^{-1} M_{\varphi(x)}^T M_{\varphi(y)} \quad (3)$$

Furthermore, the work of Cortes et al. (2007) is based on a regression technique for the purpose to learn a string to string mapping. This approach leads to many other studies in machine translation field such as the work of Wang Zhuoran et al. (2007). Wang et al. are based on a string to string mapping in order to find a linear model by using ordinary least squares (OLS) regression and n-gram string kernels on a small subset of the Europarl corpus. They use the pre-image model as a score to the standard statistical machine translation systems such as phrase-based search Koehn et al. (2003). However, this approach loses some of the main advantages of the regression approach. In fact, OLS is not necessarily the best estimator. There are some cases such as when the two (or more) of the predictor features are strongly correlated and increasing in similar way. In such cases, the determinant of the matrix $M_{\varphi(x)}^T M_{\varphi(x)}$ will be close to zero, which makes an ill-conditioned matrix.

In other words, the minimal least square estimator may causes some problems related to matrix ill-conditioning or singularity in matrix. This is caused by similar or duplicated samples that can be founded in the training set, yielding a large number of solutions. Consequently, the matrix cannot be inverted with as high precision as we'd like and the large variance affects the final parameter estimation. As an improvement of this approach, Wang and Shawe-Taylor Zhuoran and Shawe-Taylor (2008) used the L2 regularized least squares regression in machine translation.

This improvement is known as the Tikhonov regularization or ridge regression Hoerl and Kennard (1970). Even so, this solution gives preferences to a particular solution with a smaller norm by including a regularization term in this minimization as in Eq. (4):

$$min||WM_{\varphi(X)} - M_{\varphi(Y)}||_F^2 + \Gamma||W||_F^2 \quad (4)$$

However, using this regularization, the conditioning of the problem will be improved and enables a direct numerical solution. An explicit solution is given by:

$$W = (M_{\varphi(x)}^T \; M_{\varphi(x)} + \Gamma I)^{-1} M_{\varphi(x)}^T M_{\varphi(y)} \tag{5}$$

with Γ is the conditioning factor that could be determined by cross-validation and I is the identity matrix.

Although the translation quality they achieved based on Europarl corpora is still not better than statistical phrase-based (Moses framework) Koehn et al. (2007), this approach gives better results on small scale corpus.

Ergun Bicici Biçici and Yuret (2010) work is based on the feature decay FDA algorithm which is a class of instance selection algorithms. He uses feature decay in order to increase the diversity of the selected training set by devaluing the already included features Biçici and Yuret (2010). He used L1 regularized regression for sparse regression estimation of target features and graph decoding to find translation results.

The LASSO or L1-regularized method can be useful in some contexts in order to select solutions with fewer nonzero features values using this explicit formula:

$$min||W M_{\varphi(X)} - M_{\varphi(Y)}||_F^2 + \Gamma||W||_1 \tag{6}$$

Serrano et al. (2009) work is based on the learning of the translation mapping by linear regression applied to constrained hotel front desk requests domain (corpora). Once the target feature vector is obtained, they use a multi-graph search to find all possible target strings. We noticed that the majority of existing works use mainly regression or statistical techniques on spoken languages corpora. Daniel Stein work Stein (2012) and Boulares work Boulares and Jemni (2014), use statistical approach on sign language machine translation using small-sized corpora. In our previous work Boulares and Jemni (2014), we used an approach based on the two techniques of kernel regression and Statistical MT. This method requires a perfect pre-generated word-to-word alignment to give good results. The disadvantage of this work is that the perfect word-to-word alignment, cannot be generated automatically. Furthermore, the Hung-Yu Su Hung-Yu and Chung-Hsien (2009) work, relies on the extraction of the thematic relations between the grammar rules of both Chinese and Taiwanese Sign Language structure from small corpus. The extracted thematic role templates are used as Translation Memory for Statistical Machine Translation. The disadvantage of this work is that in sign language there are no general rules that can be applied automatically. Therefore, the quality of the translation depends on the quality of these extracted rules. The work of Cortes et al. (2007) aims to learn a string-to-string mapping based on the ridge regression method combined to Language Model (LM) and De Bruijn graph for the decoding process. Wang and Shawe-Taylor Zhuoran et al. (2007) work relies also on ridge regression method (L2 regularization) in order to learn the phrase-to-phrase mapping and they used the De Bruijn graph and the Language Model in the decoding process. Ergun Bicici Biçici and Yuret (2010) work is based on the use of the lasso regression method (L1 regularization) to learn the phrase-to-phrase mapping and uses the same decoding process as Wang and Shawe-Taylor Zhuoran et al. (2007). Schmidt et al. (2013) work is focused on sign language-to-text translation based on the correspondence between the mouthing and spoken language words. This work relies on a mouthing recognition system from video of a person signing into a text in a spoken language. They integrate a recognition and translation framework by adding a viseme recognizer through a lip reading system in order to optimize the recognition system and to improve the translation output. Furthermore, we observed that there are no achieved studies that rely on regression approaches in Sign language machine translation. For this purpose, we are interested by the works of Cortes et al. (2007), Wang and Shawe-Taylor Zhuoran et al. (2007), and Ergun Bicici Biçici and Yuret (2010) in order to derive benefit from statistical and regression approaches.

In this paper, we present a novel approach that consists on the use of the elastic net model for regression (well known as the combination of L1 and L2 regularization) for the purpose to learn the phrase-to-phrase mapping. For the decoding step, we rely on the multi-graph search through the De-Bruijn graph in order to find all possible target words whose mappings correspond to the translated (feature vector). In order to find the best translation from a multitude of combination (paths in the graph), we apply the Latent Semantic analysis (LSA) instead of the LM. We experimented our approach on our small-scale size ASL corpora and we obtained good results as we are presenting in Sect. 5.

3 Sign language data

Linguistic research has proven that American Sign Language has its own internal structure Neidle et al. (2000), Stokoe (1978). The grammatical structure of ASL involves the symbolic meaning of space locations and entities, known as "iconicity". In fact, iconicity occurs in spoken and gesture languages such as sign language. In sign language, the information could be interpreted by using iconic and non iconic signs. The iconic signs aim to describe an icon or a picture of some aspect of things or activities being symbolized Battison (1978). For instance, in ASL the sign "car" could be symbolized by a standard icon reflecting the word meaning (as shown in Fig. 2 a). Additionally, ASL exploits both of iconicity and space positioning (in front of the signer) for reflecting the visual aspect of the information. This specificity leads us to focus on one of the most important sign language parameters known as the locative expression of sign language.

In fact, American Sign Language use several different ways to ensure the locative reading. For example, to describe entities in a story, signers may use specific hand-shape with movement and hand orientation deployed in the signing space. This sign language manipulation is known as "the transfer of form and size". It relies on classifier predicates (CP) for symbolizing the discourse entities in the space. Indeed, classifier predicates are considered as part of lexical signs. They consist of hand-shape configuration accompanied with location, palm orientation, movement, and non-manual signals. The different relationships between entities are indicated by the movement shape and the space positioning of the appropriate classifier in the signing space. As shown in Fig. 3, the signer symbolizes the action " bites" by using the " Claw" hand-shape with the action from an entity location to another. Classifier predicates rely on several symbols for characterizing each class of entity. For example, the 3 hand-shape classifier could be used to describe several objects such as "CAR", "BOAT" and "BICYCLE". In other words, the 3 hand-shape, is a classifier symbolizing the class of objects " vehicle". The classifier (CL) with F hand-shape represents small round things: buttons, tokens, etc... The CL V (hand-shape) aims to describe legs, person walking,etc...

Furthermore, the spatial positions associated with referents can also convey locative information about the referent. For example, the phrase " the dog index" , shown in Fig. 3, could be interpreted as " the dog is there on my left". The signer add the sign "index" in order to establish a reference relation between dog and a spatial location. Signers may add a specific facial expression (e.g., spread tight lips with eye gaze to the locus) produced simultaneously with the index sign or with classifier predicates Valli and Lucas (2000). The locative expression in sign language, could be expressed by locative verbs. These verbs exploit the movement direction of the action in order to indicate the location of the entity in the space. For example, the sentence "john throw rock", the direction of the movement of the verb indicates the direction in which the object is thrown. Signers also can make reference to absolute locations, as when they use the signs for " east," " west," " north," and " south" Valli and Lucas (2000).

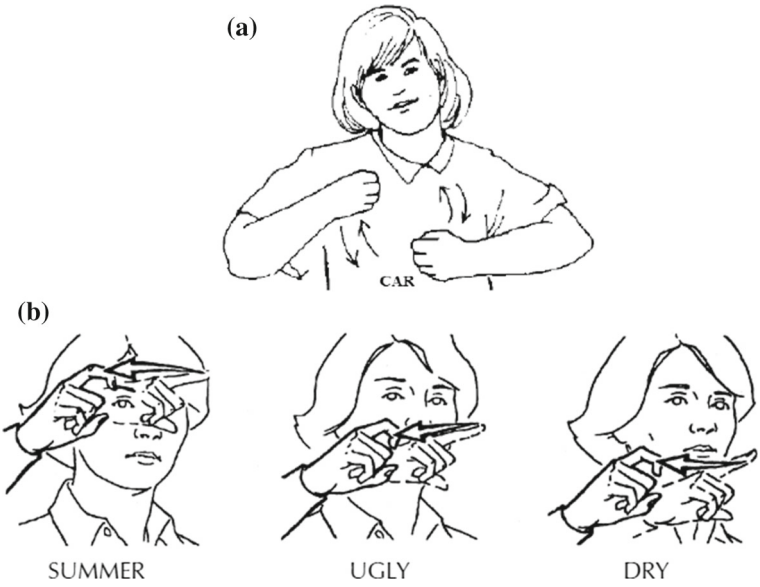


Fig. 2 **a** an overview of iconic sign "car"; **b** an example of a phonological contrast in ASL. These signs differ only in the location of their articulation

Fig. 3 An example of the sentential use of space in ASL. Nominal (cat, dog) are first associated with spatial loci through indexation. The direction of the movement of the verb (BITE) indicates the grammatical role of subject and object

Emmorey (2005) has proven that location is a part of all ASL signs and signers use location in many different ways. There are some signs that use body location such as the ASL sign "bored" which is based on the head location "nose", for sign "feel" the signer uses the chest and for sign "Russian" it is the waist. Also, signers can use the signing space surrounding them, to indicate that the sign can be in front, in left or in right. According to Sandler Sandler (1989), location is a crucial parameter that removes the semantic ambiguity of some signs such as ASL sign "summer", "ugly" and "dry" shown in Fig. 2b. All of these signs have the same signation manner and differ only in where they are articulated on the body. The study of Huenerfauth Huenerfauth and Lu (2011) has shown that signing space information improves the translation understanding. A phenomenon in which signers use special hand movements to indicate the location and movement of invisible objects (representing entities under discussion) in space around their bodies as shown in Fig. 3. In this example, the signer uses the sign "INDEX" to place the signs "DOG" and "CAT" in order to be used in the

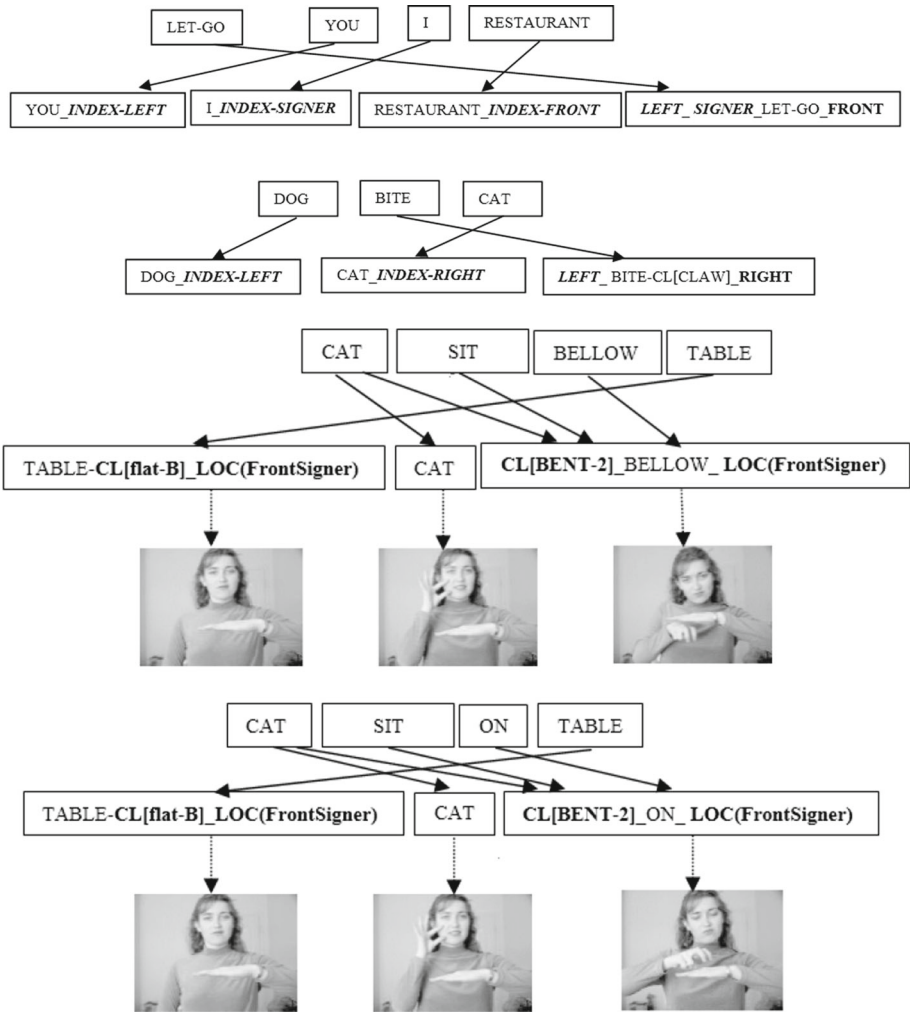


Fig. 4 An Example of the transformation of english phrase to ASL form inluding signing space information

action "BITE". Signing space information is frequent in ASL and is necessary for conveying many concepts. Therefore, the translation process that integrates spatial information is more understandable. Also, this is very helpful for deaf people who may have some difficulties in creating a mental image Charles and Rebecca (2000) reflecting the true meaning.

Furthermore, in order to be able to transcribe gesture language, we have opted to use glossing ASL to represent signs in text form. This representation differs from writing in a spoken language because when we are glossing, the target language may not have the same words order as the original language. This means that English representation needs to be translated to a glossing ASL form that includes signing space information. Consequently, this problem is suitable to be solved by machine translation technology. However, due to the lack of parallel sign language corpus that include English and glossing ASL, we have built a parallel corpus that includes 300 English text phrases and their translation using the

Table 1 Our corpus details

	Source	Target
Corpus training size	300	300
Vocabulary size	1014	966
Test set size	100	100
2-grams	557	554

glossing ASL form. We proceeded to a manual pre-treatment in order to preserve only the useful words in the English phrases (that will be translated to glossing ASL). For example, the English phrase "Let we go to the restaurant" will be transformed to "LET-GO YOU I RESTAURANT", here the words "let" and "go" are joined in order to describe one sign, the word "we" is transformed to "you i" and "to" is removed. In the same line, the sentence "The dog bites the cat" is reduced to "DOG BITE CAT". The ASL glossing form used in this work, includes the signing space information such as main entities of the initial English phrases as shown in Fig. 4.

In the examples of Fig. 4, the words order is changed according to ASL glossing Valli and Lucas (2000). For instance, "DOG BITE CAT" the entities "DOG" and "CAT" are placed on the beginning of the sentence and there is no preferential order between them ("DOG" "CAT" or "CAT" "DOG" is the same). Afterwards, the location information is added to each word using the sign "INDEX" followed by the location according to the signer (on the left, on the right, etc...). Then, these location information are used with the action (that should be placed at the end Valli and Lucas (2000)) in order to refer the pointed entities. As shown in Fig. 4, for the ASL phrase "CAT SIT BELLOW TABLE", the signer uses the passive hand with the Flat-B classifier (palm faced down) for symbolizing the table. Then he uses the dominant hand for the sign "CAT". The action "SIT BELLOW" is described by the location of the classifier BENT-2 bellow the table location. Here, the BENT-2 classifier is used to symbolize the meaning of "CAT SIT". For the ASL phrase "CAT SIT ON TABLE", the location of the BENT-2 classifier changes to be above the table location. In this example, we notice that the passive hand could be used as a location referent for the dominate hand.

Table 1 summarizes the ASL corpora information. The training size set is composed of 300 parallel sentences containing 3 to 6 words per sentence and the test set size includes 100 different phrases. The corpus phrases are composed mainly from simple SOV (Subject Object Verb) structure. In these phrases we relied on classifier predicates location, INDEX reference and absolute locations ("east", "west", "north", "south")) for describing the "locative expression" of ASL. The vocabulary size for the source phrases is 1014 and 966 for the target phrases. For the feature mapping (see Sect. 4), we generate 557 2-grams sources and 554 2-grams targets that we used to train the regression function.

4 Our approach

Our translation approach relies mainly on three steps. The first one is the feature mapping process which consists on the transformation of the input data (phrases) into feature vectors with "m" dimensional space. The purpose of this transformation is to facilitate the automatic analysis of textual data. The second step aims to learn the translation mapping between 2-grams sources and 2-grams targets. The result of this step is a linear regression function representing the translation mapping. The third and final step is the pre-image resolution that consists on determining the predicted output of an input phrase.

4.1 Feature mapping of textual data

The automatic data analysis field requires explicit feature vectors of the input data in order to derive useful information for data prediction. However, there are many cases, where the input data cannot be described by explicit feature vectors such as bio sequences, images, graphs and text documents. For such data sets, the construction of a feature extraction module can be as complex and expensive as solving the entire problem Lodhi et al. (2002). It is also possible to lose some important information during the feature extraction process. In other words, the effectiveness of a system is closely related to the accuracy and the performance of the feature extraction process. For this purpose, we may introduce kernel methods that can be considered as an efficient alternative in the feature extraction process.

The feature mapping used by Kernel methods, especially by string kernel, can be directly used by learning techniques in order to predict new data. In general, the most natural and efficient way to compare two phrases is to count the common contiguous n-gram they have. Comparing the n-spectra of two strings can give important information about their similarity, especially in the machine translation field where contiguity is a crucial parameter for the translation accuracy. For this reason, we adopted the n-spectrum weighted word kernel Shawe-Taylor and Cristianini (2004) as a feature mapping technique applied to 2-gram sequences. The feature mapping is defined by:

$$k_n(p, f) = \sum_{i=1}^{|p|-n+1} \sum_{j=1}^{|f|-n+1} k_n(p(i : i+n), f(j : j+n)) \tag{7}$$

where p represents the phrase, f is the feature and n is the length of contiguous words (grams). The computation of the n-spectrum kernel feature mapping requires $O(n|p||f|)$ operations.

4.2 The regression function

The regression technique aims to learn a model that fits data in a better way. The quality and the accuracy of unknown data prediction are crucial in the machine translation field. However, the minimal least square estimator described by formula (3) is well known as a poor prediction and interpretation model Zou and Hastie (2005), Bishop (2006). Furthermore, regularized version of OLS has been proposed in order to improve the regression accuracy such as L2-norm based regression known as ridge regression and L1-norm based regression known as Lasso regression.

Ridge regression is known as a solution to the multi-colinearity in data by adding a degree of bias to the regression estimates in order to reduce the standard errors. Multi-colinearity causes a large variance in the OLS model which makes estimations far from the true values. The existence of near-linear relationships in data causes multi-colinearity and therefore the decrease of both the accuracy of the regression coefficients and the predictability of the model. In other words, according to Zou and Hastie (2005), ridge regression cannot produce a parsimonious model. A parsimonious model ensures that the initial model will be constrained to estimate a small number of parameters. Therefore, it always keeps all the predictors in the model. Biçiçi Biçici and Yuret (2010) has shown that based on L2-norm, the obtained model cannot generate a sparse solution and the majority of the coefficients remains non-zero which makes the machine translation decoding process more complicated.

The regularization based on L1-norm proposed by Tibshirani (1996) provides a sparse model by imposing an L1-penalty on the regression coefficients. The L1-norm approach

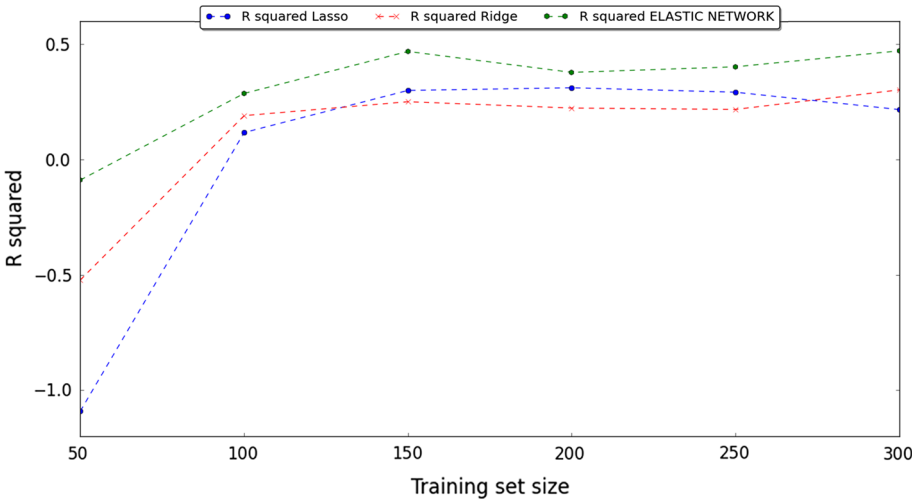


Fig. 5 R^2 values of lasso, ridge and elastic net operating on different training corpora size

proceeds to the shrinkage and automatic variable selection simultaneously in order to reduce the coefficient values. Zou and Hastie (2005) has shown that when the number of observations is greater than the number of variables, L1-norm selects at most n variables before saturation and this is an inconvenient in term of variable selection. Also, where there is a high correlated pairwise, L1-norm selects randomly only one variable. In the other hand, if the number of variables is greater than the number of observations and the predictors are highly correlated, the L2 regularization approach is more efficient compared to the L1 regularization Tibshirani (1996). Furthermore, based on the study of Tibshirani (1996), there is no uniform domination between prediction performance of the ridge and the lasso regression. Based on our numerical experience, as shown in Fig. 5, the regression performance measured by R^2 values between the lasso and the ridge regression, changes according to the corpora size. If we train the lasso and the ridge regression methods on a corpora size that varies between 50 and 100 phrases, ridge is better than lasso. However, from 150 to 250 phrases, Lasso gives better performance than ridge regression and for 300 phrases, ridge becomes again better than lasso. This numerical analysis validates the theory of the uniform domination between ridge and lasso discussed above.

Zou and Hastie (2005), Hastie et al. (2006), Trevor et al. (2009) proposed a regularization technique, called elastic net, as an improvement of the lasso technique. The elastic net approach aims to overcome the lasso problems cited above on relying on the automatic variable selection, continuous shrinkage and groups selection of correlated variables. Equation 8 is the formal description of the elastic net that contains L1 and L2 (quadratic) parts. The L1 part of the penalty generates a sparse model. The quadratic part of the penalty removes the limitation on the number of selected variables. Consequently, elastic net regularization encourages grouping effect and stabilizes the L1 regularization path. Also, real data examples show that the elastic net often outperforms the LASSO and RIDGE in terms of prediction accuracy as shown in Fig. 5. In fact, relying on our numerical analysis, we conclude that elastic net outperforms both the LASSO and the RIDGE regression in terms of prediction accuracy R^2 and Mean Squared Error (MSE) despite the corpora size variation (see Figs. 5, 6). For this reason, we adopted elastic net as a regression function for our machine translation process.

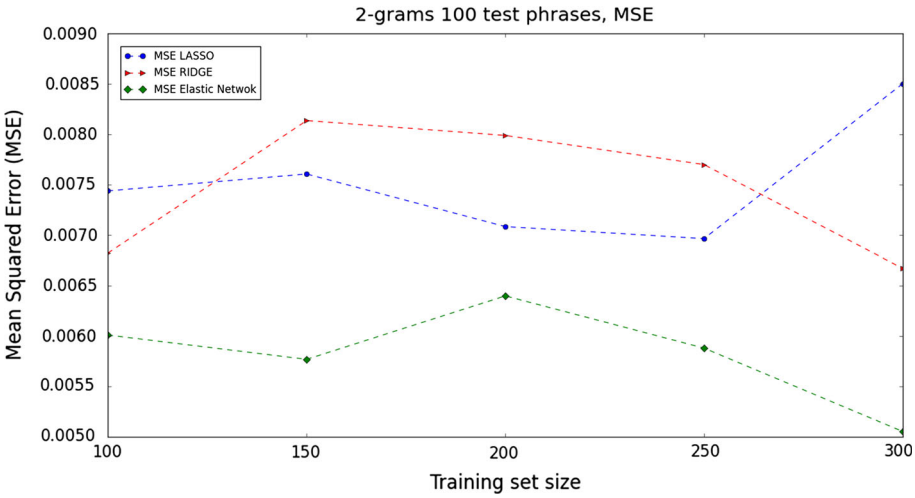


Fig. 6 MSE comparing between Lasso, Ridge and Elastic net regression function with training set size from 100 to 300 phrases

The elastic net technique is based mainly on the combination of the L1 and L2 penalties:

$$L(\Gamma_1, \Gamma_2, W) = ||WM_{\varphi(X)} - M_{\varphi(Y)}||^2 + \Gamma_2||W||^2 + \Gamma_1||W||_1 \tag{8}$$

where $||W||^2 = \sum_{j=1}^{f} W_j^2$, $||W||_1 = \sum_{j=1}^{f} ||W_j||$, f: features with $j = 1...f$. With the minimization of Eq.(8), the naive elastic net estimator becomes:

$$\hat{W} = argmin_w L_{1_{ratio}}(\Gamma_1, \Gamma_2, W) \tag{9}$$

with $L_{1_{ratio}} = \frac{\Gamma_2}{\Gamma_1+\Gamma_2}$. Solving $\hat{W}$ in Eq.(8) is equivalent to the optimization problem:

$$\hat{W} = argmin_w ||WM_{\varphi(X)} - M_{\varphi(Y)}||^2 + (1 - L_{1_{ratio}})||W||_1 + L_{1_{ratio}}||W||^2 \tag{10}$$

with $(1 - L_{1_{ratio}})||W||_1$, $L_{1_{ratio}}||W||^2$: are the two parameters respectively lasso and ridge penalty to form the elastic net penalties. When $L_{1_{ratio}} = 1$ the estimator becomes ridge. When $L_{1_{ratio}} = 0$ the estimator becomes lasso and if $0 < L_{1_{ratio}} < 1$ the penalty is a combination of $L_1$and L_2. As mentioned in Zou and Hastie (2005), the naive version of elastic net method finds first the ridge regression coefficients by fixing the Γ_2 and then performs a LASSO shrinkage to generate a sparse model. The quadratic part of the penalty removes the limitation on the number of selected variables, leads to a grouping effect and stabilizes the l1 regularization path. In Zou and Hastie (2005), the authors, presented a solution to solve the naive elastic net problem efficiently:

$$\hat{W} = argmin_W W^T \left(\frac{M_{\varphi(X)}^T M_{\varphi(X)} + \Gamma_2 I}{1 + \Gamma_2} \right) W - 2M_{\varphi(Y)}^T M_{\varphi(x)} W + \Gamma_1||W||_1 \tag{11}$$

4.3 The decoding Problem

4.3.1 De Bruijn graph

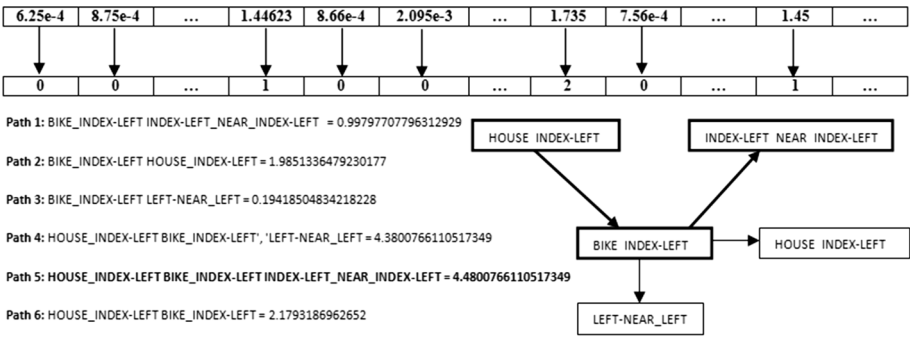
The decoding problem, known as the pre-image problem, aims to find the target sentence Y from the feature vector $\varphi_{(y)}$ predicted in Eq. (1). In fact, we are based on Eq. (11) to predict the $\hat{W}$ estimator that will be used to find the new feature values through Eq.(12):

$$Y = \varphi_{(y)}^{-1} = argmin||\hat{W}\varphi_{(x)} - \varphi_{(y)}||^2 \tag{12}$$

The obtained vector values are rounded in order to to obtain integer count as in Cortes et al. (2007). Thus, the pre-image solution is achieved by building a De Bruijn graph with the non zero features values in order to connect the 2-gram features in a same graph. This technique aims to seek all possible paths between nodes and therefore, it identifies all the possible translations as shown in Fig. 7.

4.3.2 LSA search

Latent Semantic Analysis (LSA) is a statistical technique that aims to discover hidden concepts in order to extract relationships between documents. Each document and term (word) is represented by a vector with elements that expose document to document similarities or semantic relationship. Each element in the vector represents the degree of participation of the document or term in the corresponding concept. LSA was presented as an information retrieval method and sometimes called Latent Semantic Indexing (LSI) Deerwester et al. (1990). In Biçici and Yuret (2010) and Zhuoran et al. (2007) works, the automatic translation selection process was based on language model to select the most appropriate translation between a set of possible translations generated from the de-bruijn graph. In fact, as shown in Sect. 5, using LSA in the pre-image step precisely in the translation selection process will improve the translation result in term of accuracy and quality. This improvement is thanks to the semantic aspect of the LSA technique allowing the description of the semantic similarities between the different possible translations generated from the de-bruijn graph and



	Phrase 1	Phrase 2	Phrase 3	…	Phrase n
Path1	0.126685	**0.9979770779631292 9**	0.1342998	…	0.76990881
Path2	0.74555321	1.094432	**1.9851336479230177**	…	1.309332165
Path3	0.002344453	0.012995	**0.19418504834218228**	…	0.000435541
Path4	1.654908	**4.3800766110517349**	2.12567190	…	0.548832109
Path5	**4.4800766110517349**	2.1980743	1.756322854	…	3.212987765
Path6	2.1298547	0.981184532	1.45210998	…	**2.1793186962652**

Fig. 7 An overview of our pre-image solution

those in the corpora. In other words, our goal is to find the most similar translation between the de-bruijn set of translations and our corpus translations throw the following steps :

Step 1: In this step, we create a vector for each the target phrases in our corpora with all the 2-grams terms based on formula (7) in order to obtain $n * m$ document-term matrix. Formally let A be the $n * m$ document-term matrix of the documents collection. Each column of A corresponds to 2-gram term. The dimensions of A, m and n, correspond respectively to the number of words and documents in the collection. we apply formula (7) for weighting all the elements of the matrix.

Step 2: We perform a dimension reduction through Singular Value Decomposition (SVD) De Lathauwer et al. (2000) on A as follow:

$$M = USV^T \tag{13}$$

where $U^T U = I$, $V^T V = I$; the columns of U are orthonormal eigenvectors of AA^T , the columns of V are orthonormal eigenvectors of $A^T A$, and S is a diagonal matrix containing the square roots of eigenvalues from U or V in descending order.

Step 3: This step aims to find the most similar translation generated by the de-bruijn graph. For each unknown translation T generated from the de-bruijn paths, we apply step 1 on T in order to obtain the vector K. We perform an SVD with the same dimension reduction parameter on K to obtain K'. For the purpose to extract the most similar translation to our target corpus phrases, we use the following formula:

$$Tr_{phrase} = argmax(K' M^T) \tag{14}$$

We repeat step 3 for each generated de-bruijn translation in order to obtain a vector of Tr_{phrase} formulated by $Vect_{Tr_{phrase}}$. By applying the formula (15), we obtain the most similar and appropriate translation as shown in Fig. 7:

$$Phrase_{Tr} = argmax(Vect_{Tr_{phrase}}) \tag{15}$$

5 Experimental study

In order to validate our approach we conducted an experimental study. As mentioned above, we built a corpus of 300 parallel phrases and we use it as a framework of our experiments. First, we applied the three techniques Ridge, Lasso and Elastic Net on our corpus and then, we compared the results. We analyzed ,in particular, the regression function results and the MT evaluation reports. Furthermore, we used in our experiments both LM and LSA in the decoding process. In the next two sub sections, we present the main findings we obtained.

5.1 The regression function

We conducted a detailed experimental study concerning the Elastic Net regression function used in our solution. Figures 8 and 9 represent respectively an experimental comparison of the Elastic net fitting accuracy in term of R^2 and MSE with different values of L1 ratio and training set size from 100 to 300 phrases. It is obvious to notice that the variation of the corpus training set size and the L1 ratio penalty (that reflects the degree of the combination of L1 and L2 norm in the elastic net regression function) affects the R^2 fitting accuracy.

As shown in Fig. 8, for 100 phrases training size the L1_ratio = 0.1 has the best fitting value (R2) comparing to L1_ratio from 0.2 to 0.9. Also, from Fig. 9, we notice that for 100 phrases training size the L1_ratio = 0.1 has the minimum Mean squared error (MSE). For

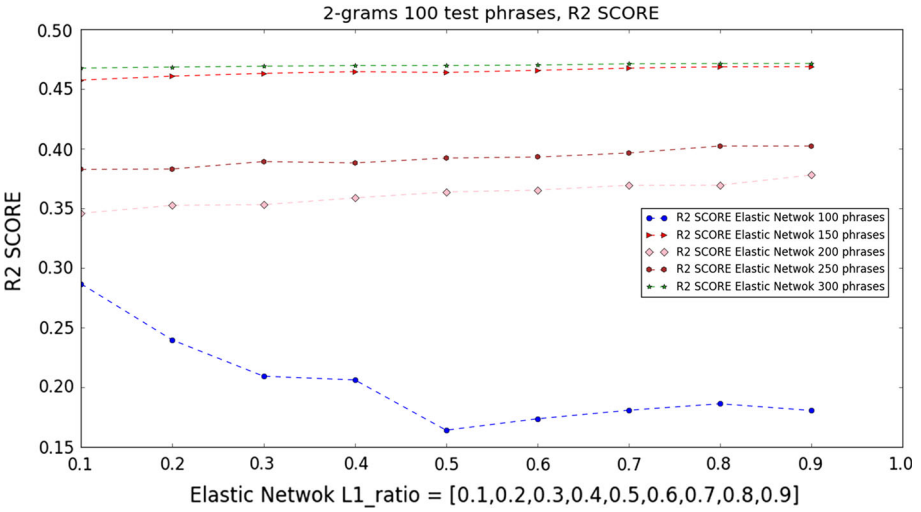


Fig. 8 Experimental comparing of Elastic net fitting accuracy in term of R^2 with different values of L1 ratio and training set size

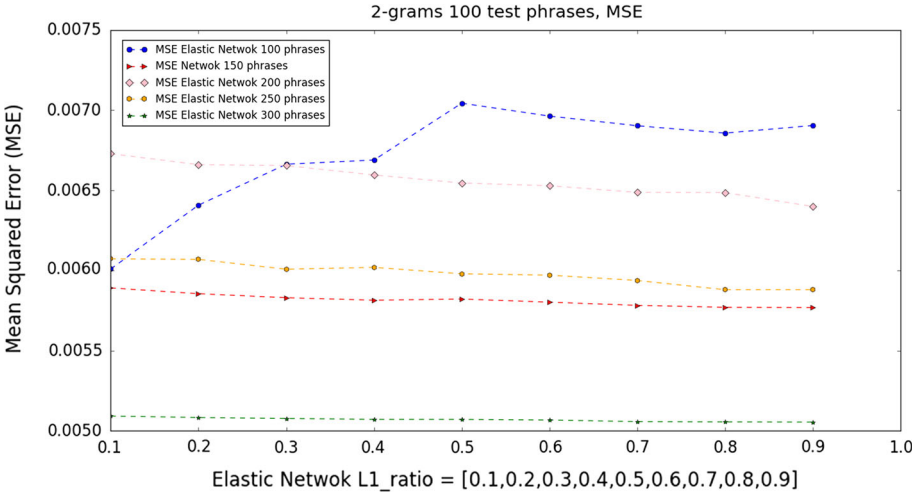


Fig. 9 Experimental comparing of Elastic net fitting accuracy in term of Mean squared error (MSE) with different values of L1 ratio and training set size

300 phrases the L1_ratio = 0.9 has the highest fitting value (R^2score) and the minimum MSE. With 100 training phrases and for :

- $L1_{ratio_{Elastic}}(0.1)$, $R^2_{Elastic}(0.28) > R^2_{Elastic}(0.19) > R^2_{Elastic}(0.11)$.
- $MSE_{Elastic}(0.0060) < MSE_{Ridge}(0.0068) < MSE_{Lasso}(0.0074)$.

From these experiments, we deduce that Elastic Network regression function improves the performance of the training data adjustment and this is thanks to the combination of advantages of both Lasso and Ridge regression functions.

5.2 Machine translation evaluation

We remind that we used a corpus that contains 300 parallel phrases from English text to AS (including signing space). We used as well, 100 different testing parallel phrases. As mentioned in section 3, the reduced size of our corpus is due to the lack of sign language corpora.

As shown in Fig. 10, our experimental results show that the variation of the bleu score is according to the variation of two parameters: the LSA dimension used in the translation selection process and the $L1_{ratio}$ of the Elastic Network Regression function. Furthermore, it is clear that the highest $BLEU$ score is obtained when $LSA_{dimension} = 250$. This LSA dimension value (250) is the highest possible dimension that can be used experimentally for 300 phrases corpus size and this is based on the sklearn python library ($sklearn\ SingleValueDecomposition\ SVD$).

Furthermore, Figs. 11, 12, show also that based on $LSA_{dimension} = 250$ in the translation selection process, our Machine Translation approach has the highest *METEOR* Banerjee and Lavie (2005) and *NIST* Doddington (2002), Przybocki (2004) scores.

We conducted also an experiment based on the corpus size variation, and as shown in Fig. 13 and Tables 2, 3 , besides corpus size = 100, our solution performs better than Zhuoran wang approach (2008) Zhuoran and Shawe-Taylor (2008) and E. Biçiçi Biçici and Yuret (2010) respectively in term of the metrics : BLEU, METEOR, NIST and F1-MESURE scores. It is also clear (in Fig. 13) that from 65 to 130 phrases the bleu score of wang approach is higher than our bleu score. The performance of our approach becomes stable and better than wang approach, when the size of the corpus reaches and exceeds 130 phrases.

As shown in Fig. 13:

- 50 phrases : $Bleu_{enet\&lsa}(6.52) > Bleu_{wang}(6.04) > Bleu_{Biçiçi}(5.93)$.
- From 50 to 65 phrases : $Bleu_{enet\&lsa} > Bleu_{wang} > Bleu_{Biçiçi}$.
- From 65 to 130 phrases : $Bleu_{wang} > Bleu_{enet\&lsa} > Bleu_{Biçiçi}$.
- From 130 to 300 phrases : $Bleu_{enet\&lsa} > Bleu_{wang} > Bleu_{Biçiçi}$.

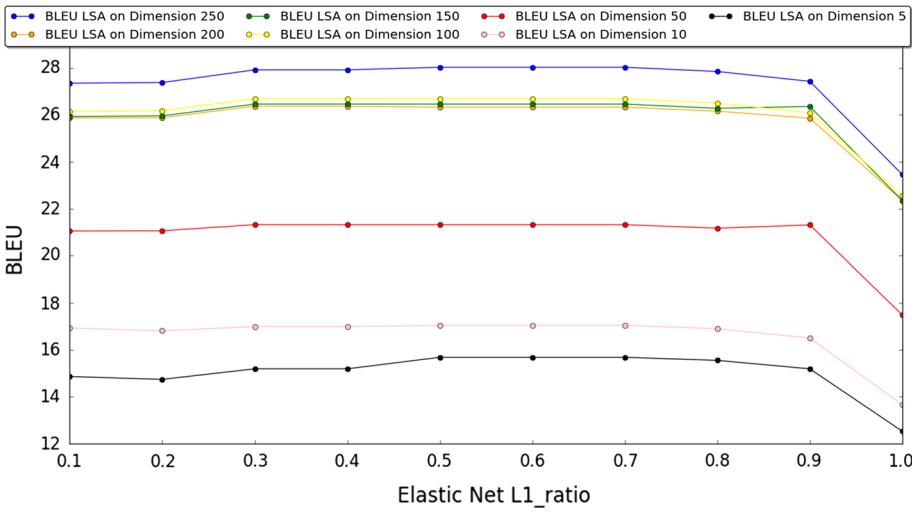


Fig. 10 BLEU scores using Elastic Net & LSA search with the variation of LSA dimension and $L1_{ratio}$

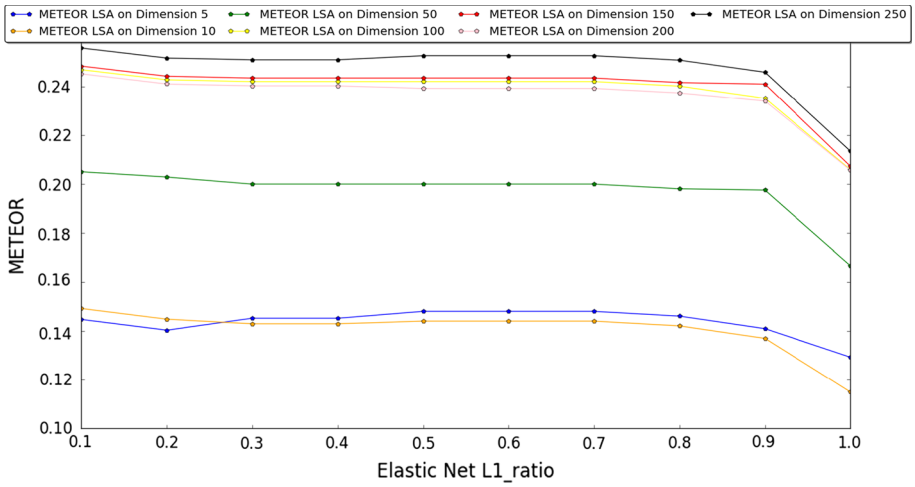


Fig. 11 METEOR scores using Elastic Net & LSA search with the variation of LSA dimension and $L1_{ratio}$

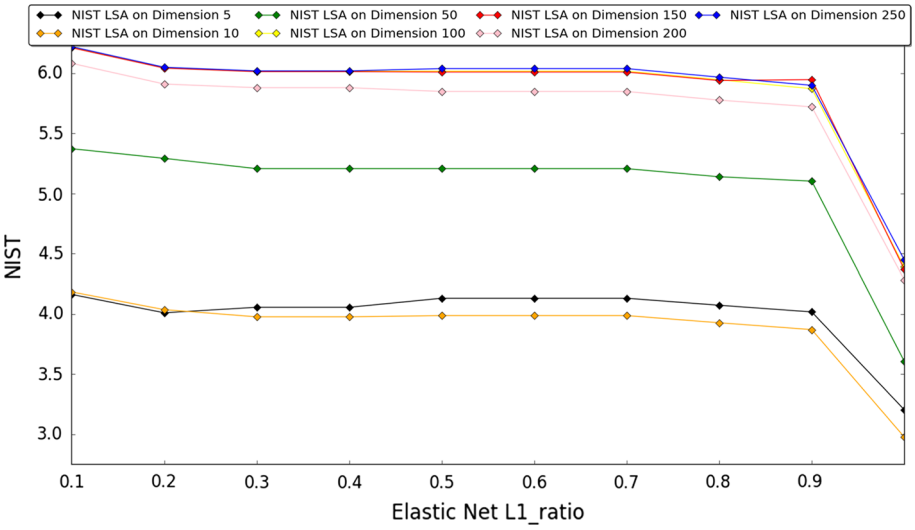


Fig. 12 NIST scores using Elastic Net & LSA search with the variation of LSA dimension and $L1_{ratio}$

– 150 phrases : $Bleu_{enet\&lsa}(18.75) > Bleu_{wang}(17.38) > Bleu_{Biçiçi}(15.56)$.
– 250 phrases : $Bleu_{enet\&lsa}(24.05) > Bleu_{wang}(20.22) > Bleu_{Biçiçi}(16.71)$.

We experimented also Lasso, Ridge ans elastic net regression methods with two different decoding process. In the first decoding process, we used the classical language model with de bruijn graph and for the second, we used LSA as decision rule with de bruijn graph.

As shown in Table 4, we compared the MT evaluation report of Elastic Net with LM decoding and Elastic Net with LSA decoding and we obtained best translation results (in term of the metrics BLEU, NIST, METEOR and F1-MEASURE) using the LSA as a decision rule with the de bruijn graph. Also, we compared the MT evaluation report of Lasso, Ridge and Enet with the use of both LM and LSA in the decoding process (see Fig. 14, Table 5). We

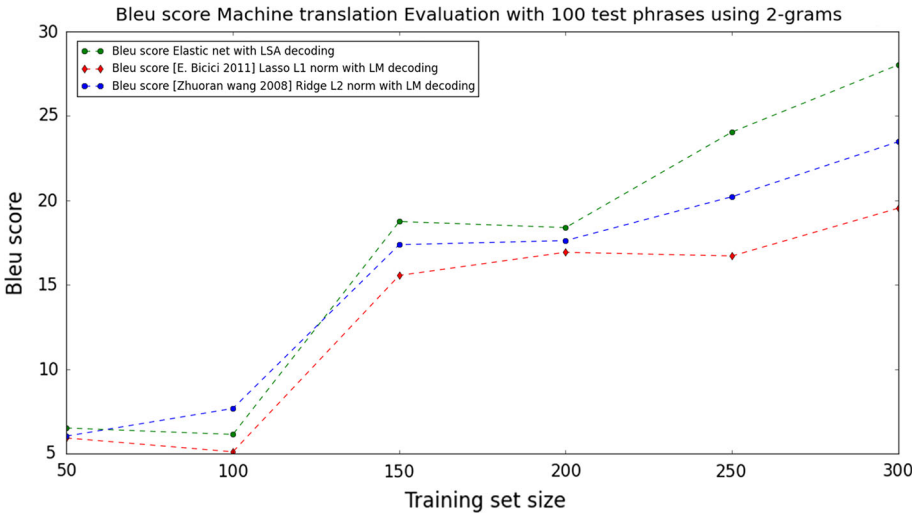


Fig. 13 BLEU scores comparison of Zhuoran et al. (2007), Biçici and Yuret (2010) and our approach based on our corpus size variation from 50 to 300 phrases using 2-grams

Table 2 A comparison of Machine Translation approches based on MT Evaluation methods experimented on 2-grams and 50 to 250 phrases corpus

Corpus size	METEOR	F1-MEASURE	NIST
	50		
Biçiçi Lasso	0.1098	0.2543	1.2336
WANG Ridge	0.1127	0.2637	1.5965
Elastic Net & LSA search	0.1191	0.2782	1.7909
	100		
Biçiçi Lasso	0.0900	0.2025	0.6316
WANG Ridge	0.1270	0.2888	2.3120
Elastic Net & LSA search	0.1066	0.2404	1.0897
	150		
Biçiçi Lasso	0.1687	0.3610	3.0350
WANG Ridge	0.1821	0.3986	4.2108
Elastic Net & LSA search	0.2078	0.4573	5.0046
	200		
Biçiçi Lasso	0.1765	0.3761	3.3192
WANG Ridge	0.1881	0.4076	4.4930
Elastic Net & LSA search	0.2107	0.4633	5.3645
	250		
Biçiçi Lasso	0.1777	0.3811	3.8505
WANG Ridge	0.2099	0.4591	5.4423
Elastic Net & LSA search	0.2465	0.5298	6.4515

Table 3 A comparison of Machine Translation approches based on MT Evaluation methods experimented on 2-grams and 300 phrases corpus

	BLEU	METEOR	F1-MEASURE	NIST
Biçiçi Lasso	19.54	0.2030	0.4376	4.5339
WANG Ridge	23.48	0.2321	0.5036	5.8507
MOSES	16.28	0.2425	0.5140	5.1468
Elastic Net & LSA	28.03	0.2527	0.5495	6.0384

Table 4 Experimental comparing of Elastic Net based MT using both Language Model (LM) and LSA-based decision rule

Size	BLEU		NIST		METEOR		F1-Mesure	
	LM	LSA	LM	LSA	LM	LSA	LM	LSA
50	6.04	**6.52**	1.596	**1.790**	0.112	**0.119**	0.263	**0.278**
100	5.52	**6.14**	0.685	**1.089**	0.094	**0.106**	0.211	**0.240**
150	17.47	**18.75**	4.123	**5.004**	0.186	**0.207**	0.410	**0.457**
200	17.51	**18.39**	4.591	**5.318**	0.191	**0.210**	0.417	**0.463**
250	20.16	**24.05**	5.506	**6.451**	0.211	**0.246**	0.464	**0.529**
300	23.00	**28.03**	5.791	**6.038**	0.232	**0.252**	0.506	**0.549**

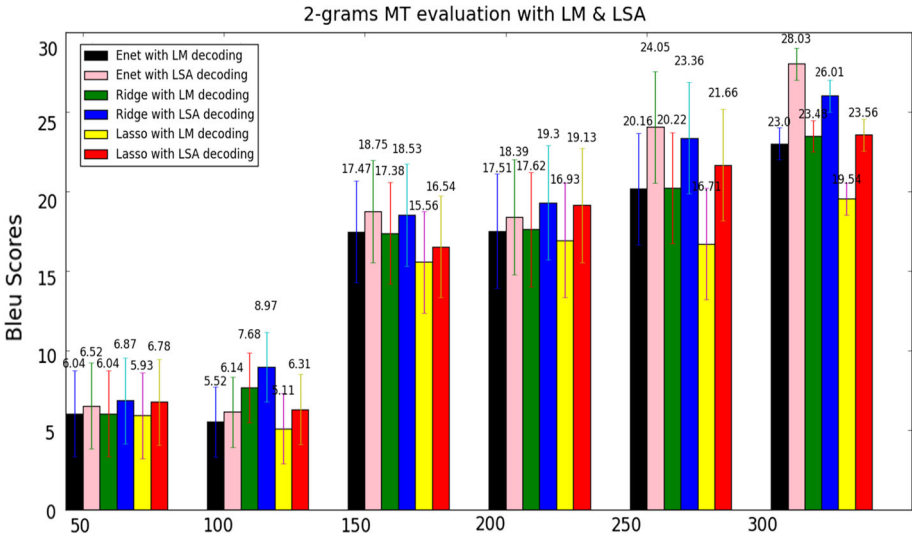


Fig. 14 A comparison of Bleu scores of Ridge, Lasso and Enet with LM and LSA decoding

obtained an improvement in term of the metrics : Bleu, Meteor, F1-MEsure and Nist scores of the translation results of Lasso, Ridge and Enet using LSA compared to LM decision rule (see Fig. 14, Table 5). In other words, the translation quality and accuracy using LSA with the de-bruijn graph in the decoding process is more powerful than the use of the classical LM with de-bruijn. This improvement is explained by the requirement of very big training

Table 5 MT evaluation report of Lasso, Ridge and Enet with both LM and LSA decoding

		METEOR			F1-MESURE			NIST		
		Enet	Lasso	Ridge	Enet	Lasso	Ridge	Enet	Lasso	Ridge
50	LM	0.1127	0.1098	0.1127	0.2637	0.2543	0.2637	1.5965	1.2336	1.5965
	LSA	0.1191	0.1197	0.1216	0.2782	0.2744	0.2811	1.7909	1.4813	1.8005
100	LM	0.0948	0.0900	0.1270	0.2115	0.2025	0.2888	0.6851	0.6316	2.3120
	LSA	0.1066	0.1053	0.1423	0.2404	0.2342	0.3226	1.0897	1.0350	2.8551
150	LM	0.1869	0.1687	0.1821	0.4106	0.3610	0.3986	4.1236	3.0350	4.2108
	LSA	0.2078	0.1774	0.1998	0.4573	0.3827	0.4395	5.0046	3.5741	4.9285
200	LM	0.1916	0.1765	0.1871	0.4173	0.3761	0.4076	4.5912	3.3192	4.4930
	LSA	0.2107	0.2024	0.2127	0.4633	0.4331	0.4616	5.3180	4.2488	5.3953
250	LM	0.2118	0.1777	0.2099	0.4640	0.3811	0.4591	5.5067	3.8505	5.4423
	LSA	0.2465	0.2128	0.2375	0.5298	0.4543	0.5136	6.4515	5.0309	6.3126
300	LM	0.2325	0.2030	0.2321	0.5069	0.4376	0.5036	5.7913	4.5339	5.8507
	LSA	0.2558	0.2243	0.2554	0.5495	0.4789	0.5458	6.0384	5.4383	6.4880

data size (contiguous word (2-grams) frequency) for the 2-grams language model to reach the optimal performance. By cons, LSA is based on the principal component decomposition method which has a less degree of influence to small training data size compared to LM method and this is proved experimentally.

In summary, based on our ASL corpus (300 phrases), we tested the performance of our machine translation approach comparing to Zhuoran et al. (2007), Biçici and Yuret (2010) and Koehn et al. (2007) works. We concluded that our approach improves the translation accuracy and quality in term of BLEU Papineni et al. (2002), Denoual and Lepage (2005), NIST, METEOR and F1-Measure Powers (2011), Lavie et al. (2004) scores (see Table 2). We noticed that the performance of our machine translation approach in term of the previous metrics Coughlin (2003), Doddington (2002), Finch et al. (2005), becomes stable from 130 phrases (as shown in Fig. 13; Tables 2, 3). For 300 parallel phrases corpus and for $L1_{ratio} = 0.5$ with $LSA_{dimension} = 250$, we achieved good results (as shown in Table 3) with $BLEU = 28.03$, $METEOR = 0.2527$, $NIST = 6.0384$ and $F1-Measure = 0.5495$. Finally and according to these different experimentation and with the corpus we used, we conclude that our approach gives in general better results than Bicici Lasso, Wang Ridge and phrase-based MT.

6 Conclusion

In this paper, we presented a novel approach for sign language machine translation based on the Elastic network regression & LSA search. We created a sign language parallel corpus containing 300 phrases including signing space specificity. We used the n-spectrum weighted word kernel as a feature mapping technique applied to 2-gram sequences. In the training process, we proved that the Elastic Net regression function performs better than LASSO and RIDGE regression functions. In the other hand, we used the de-bruijn graph combined to the Latent Semantic Analysis approach in order to improve the LM decoding process, and therefore, to improve the translation quality and accuracy. The experimental study confirmed the advantage of our approach as we obtained good results in term of the well known met-

rics : BLEU, METEOR, NIST and F1-MEASURE. Generally, our approach leads to good results when applied to a reduced domain with simple ASL Subject Object Verb form (short phrases length). The ultimate goal of our future research is to increase the size of our sign language corpus in order to cover all possible ASL vocabulary and to improve consequently the translation quality.

References

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, June 2005

Battison R (1978) Lexical borrowing in American sign language: phonological and morphological restructuring. Linstok Press, Silver Spring

Biçici E, Yuret D (2010) L1 regularization for learning word alignments in sparse feature matrices. In: Proceedings of the Computer Science Student Workshop, 2010

Bishop CM (2006). Pattern recognition and machine learning. Springer ISBN 978-0-387-31073-2, 2006

Boulares M, Jemni M (2012) Mobile sign language translation system for deaf community. In: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility. ACM, ISBN: 978-1-4503-1019-2

Boulares M, Jemni M (2014) Combined methodology based on kernel regression and kernel density estimation for sign language machine translation. Advances in neural networks. ISNN'14, Springer pp 374–384

Charles A, Rebecca S (2000) Reading optimally builds on spoken language implication for deaf readers. Learning research and development center University of Pittsburgh, Pittsburgh

Colin CA et al (1997) An R-squared measure of goodness of fit for some common nonlinear regression models. J Econ 77(2):1790–1792. doi:10.1016/S0304-4076(96)01818-0 (PMID 11230695)

Cortes C, Mehryar M, Jason W (2007) A general regression framework for learning string-to-string mappings. In: Predicting structured data. The MIT Press, pp 143–168, September 2007

Coughlin D (2003) Correlating automated and human assessments of machine translation quality. MT Summit IX, New Orleans

De Lathauwer L et al (2000) A multilinear singular value decomposition. SIAM J Matrix Anal Appl 21(4):1253–1278

Deerwester SC et al (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci (1986–1998); Sep 1990; 41, 6; ABI/INFORM Global pg. 391

Denoual E, Lepage Y (2005) BLEU in characters: towards automatic MT evaluation in languages without word delimiters. In: Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing, pp 81–86

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings of the Human Language Technology Conference (HLT), San Diego, pp 128–132

Doddington G (2002) The NIST automated measure and its relation to IBMs BLEU. In: Proceedings of LREC-2002 Workshopon Machine Translation Evaluation: Human Evaluators Meet Automated Metrics

Efthimiou E et al (2009) Sign language recognition, generation, and modelling: a research effort with applications in deaf communication. In: Universal access in human–computer interaction addressing diversity. Springer, pp 21–30

Efthimiou E et al (2007) GSLC: creation and annotation of a greek sign language corpus for HCI. In: Stephanidis C (ed) HCI 2007, LNCS, vol 4554. Springer, Heidelberg, pp 657–666

Emmorey K (2005) The confluence of space and language in signed languages. Linguist Am Sign Lang Introd 3:318–346

Finch A, Hwang Y-S, Sumita E (2005) Using machine translation evaluation techniques to determine sentence-level semantic equivalence. IWP2005 2005

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. doi:10.2307/1267351 JSTOR 1271436

Huenerfauth M, Lu P (2011) Effect of spatial reference and verb inflection on usability of sign language animations. Springer-Verlag Univ Access Inf Soc. doi:10.1007/s10209-011-0247-7

Hung-Yu S, Chung-Hsien W (2009) Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory. IEEE Trans Audio Speech Lang Process 17(7):1305–1315 September 2009

Koehn P et al (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
Koehn P, Hoang H (2007) Factored translation models. In: Proceedings of EMNLP-CoNLL'07
Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of HAACL-HLT'03, pp 48–54
Lavie A, Sagae K, Jayaraman S (2004) The Significance of Recall in Automatic Metrics for MT Evaluation. In: Proceedings of AMTA 2004, Washington DC. September 2004
Leslie C, Eskin E, Stafford W (2002) The spectrum kernel: a string kernel forsvm protein classifi-cation. Pacific symposium on Biocomputing. pp 566–575. 2002
Lodhi H, Saunders C, Shawe-Taylor J, Nello C, Watkins C (2002) Text classification using string kernels. J Mach Learn Res 2:419–444
Neidle C et al (2000) The syntax of American sign language: functional categories and hierarchical structure. MIT Press, Cambridge
Neidle C, Sclaroff S (2002) Data collected at the national center for sign language and gesture resources. Boston University, Boston
Papineni K et al (2002) BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp 311–318. CiteSeerX: 10.1.1.19.9416
Powers DMW (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2(1):37–63
Przybocki M (2004) NIST machine translation 2004 evaluation summary of results. In: Machine Translation Evaluation Workshop
Sandler W (1989) Phonological representation of the sign: linearity and nonlinearity in American sign language. Foris, Dordrecht
Schmidt C, Koller O, Ney H, Hoyoux T, Piater J (2013) Using viseme recognition to improve a sign language translation system. International Workshop on Spoken Language Translation, pp 197–203
Serrano N, Andres-Ferrer J, Casacuberta F (2009) On a kernel regression approach to machine translation. In: Iberian Conference on Pattern Recognition and Image Analysis, pp 394–401, 2009
Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge university press, Cambridge
Stein D (2012) Analysis, preparation, and optimization of statistical sign language machine translation. Mach Trans 26(4):325–357
Stokoe WC (1978) Sign language structure. 1978. ERIC
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc. Series B (Methodol). 267–288
Trevor H, Jonathan T, Robert T, Guenther W (2006) Forward stagewise regression and the monotone lasso. Electron J Stat 1:1–29
Trevor H, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, New York
Valli C, Lucas C (2000) Linguistics of American sign language: an introduction. Gallaudet University Press, Washington, DC
Watkins C (2000) Dynamic alignment kernels. Adv Large Margin Classif, 39–50
Zhuoran W, Shawe-Taylor J (2008) Kernel regression framework for machine translation: UCL system description for WMT 2008 shared translation task. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp 155–158, 2008
Zhuoran W, Shawe-Taylor J, Sandor S (2007) Kernel regression based machine translation. In: Human Language Technologies. The Conference of the North American Chapter of the Association for Computational Linguistics; pp 185–188, 2007
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc: Series B (Stat Methodol) 67(2):301–320 April 2005