Gupta, Parth , Costa-jussà, Marta R. , Rosso, Paolo, Banchs, Rafael E. (2016) A Deep source-context feature for lexical selection in statistical machine translation. Pattern Recognition Letters, Vol. 75, May 2016, pàg. 24-29. Doi:10.1016/j.patrec.2016.02.014.

Pattern Recognition Letters
journal homepage: www.elsevier.com

A Deep Source-Context Feature for Lexical Selection in Statistical Machine Translation

Parth Gupta[a], Marta R. Costa-jussà[b,**], Paolo Rosso[a], Rafael E. Banchs[c]

[a] *PRHLT Research Center, Universitat Politècnica de València*
[b] *TALP Research Center, Universitat Politècnica de Catalunya, Barcelona*
Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico
[c] *Human Language Technology, Institute for Infocomm Research, Singapore*

ABSTRACT

This paper presents a methodology to address lexical disambiguation in a standard phrase-based statistical machine translation system. Similarity among source contexts is used to select appropriate translation units. The information is introduced as a novel feature of the phrase-based model and it is used to select the translation units extracted from the training sentence more similar to the sentence to translate. The similarity is computed through a deep autoencoder representation, which allows to obtain effective low-dimensional embedding of data and statistically significant BLEU score improvements on two different tasks (English-to-Spanish and English-to-Hindi).

1. Introduction

Source context is usually very relevant when translating texts. However, standard phrase-based statistical machine translation (SMT) systems use a source context limited to the words that compose the translation units. The source-context information becomes specially necessary when translating from different domains. Also, the source-context information is important when the source language has source words with the same form (spelling) that can be translated into a different form target words.

Addressing the two different motivations, source context information has been introduced in the phrase-based system from different perspectives: lexical semantics or topic adaptation (Section 2). The former uses different classification techniques to decide the meaning of words with multiple translations. The latter explore different topic feature functions.

In this paper, we propose to enhance the context-awareness of translation units by taking into account the semantic context provided by the source sentence to be translated (Section 3). This allows to introduce a new feature function for each translation unit that informs about the similarity of the input sentence to be translated with the source sentence from which the translation unit was extracted from. The methodology proposed and evaluated in this work is based on the source context similarity approach presented in (Banchs and Costa-jussà, 2011) that use latent semantic analysis (LSA) to compute similarity among different contexts. Different from that work, we introduce the use of auto-encoders to construct a deep representation of sentences in a reduced space before computing similarities among sentences (used as source context of the translation units). Our algorithm was tested in the international evaluation of the Workshop on Statistical Machine Translation 2014 (Costa-jussà et al., 2014). We evaluate the use of features learned by deep autoencoders as the modelling framework for assessing semantic similarity among sentences (Section 4). Deep learning has shown to outperform compared to other generative models like the already mentioned LSA (Hinton and Salakhutdinov, 2006) and LDA (Salakhutdinov and Hinton, 2009; Srivastava et al., 2013). After the introduction of the unsupervised pretraining (Hinton and Salakhutdinov, 2006; Erhan et al., 2010), deep autoencoders, used in this work to estimate similarity between sentences in the contextual latent space, can efficiently be trained. Deep learning algorithms implemented using GPUs are highly scalable. Domain adaptation for already trained model, which is an important issue of contextual similarity methods, can very effectively be handled with deep learning methods Glorot et al. (2011); Bengio (2012). For similar methods like LSA, the context matrix has to be factorized from the scratch for the adaptation. With this methodology,

[**] Corresponding author
e-mail: `marta.ruiz@upc.edu` (Marta R. Costa-jussà)

the goal is to improve the translation output in terms of lexical selection.

Experiments on standard data collections for English-Spanish and English-Hindi translation tasks show the proposed method performs significantly (statisticallly) better than the baselines (Section 5). We also present a thorough analysis and scalability aspects of the proposed method.

The rest of the paper is organized as follows. Section 2 reports an overview of the related work on introducing source context information and using deep learning in standard SMT systems. Section 3 presents how the phrase-based model is extended with source context information. Section 4 explains the deep representation of sentences, which is used to better compute similarities among source contexts. Section 5 describes the experiments where we proof the relevance of the technique and section 6 concludes.

2. Related Work

Since the main novelty of this paper is adding source context knowledge by means of deep learning techniques in a standard phrase-based SMT system, we give an overview of some relevant works (without aiming at completeness) in this area.

2.1. Adding source context in SMT

As mentioned in the previous Section, addressing the two different motivations, source context information has been introduced in the phrase-based system from different perspectives: lexical semantics or topic adaptation.

As lexical semantics works, Carpuat and Wu (2005) introduce word sense disambiguation techniques. Bonet et al. (2009) train local classifiers using linguistic and context information to translate a phrase. Haque (2010) use different syntactic and lexical features which are proposed for incorporating information about the neighbouring words and report a complete state-of-the-art on introducing source context in a phrase-based system that the reader can refer to.

From the topic adaptation perspective, works basically focus on addressing the challenge of translating in different domains. For example, Banchs and Costa-jussà (2011) use latent semantic analysis (LSA) to compute similarity among different contexts. More recently, Chen et al. (2013) compute phrase pair features from vector space representations that capture domain similarity to a development. Hasler et al. (2014) use latent Dirichlet allocation (LDA) to compute topic feature functions.

2.2. Using Deep Learning techniques in SMT

For the last 10 years, there has been an increase of studies on MT that use different strategies based on deep learning. What is worth noticing is that there has been a huge explosion of works on this topic in the last big conferences of ACL, NAACL and EMNLP. Most of the approaches try to modify some feature or model from an standard SMT system. Other few works propose novel MT architectures.

First works in adding deep learning in SMT systems are those that use continuous-space or neural language models, e.g. Schwenk et al. (2006); Vaswani et al. (2013). Other ones smooth bilingual language models inspired on the previous ones, e.g. Schwenk et al. (2007); Zamora-Martínez et al. (2010) After that, Liu et al. (2013) use deep learning algorithms to improve translation and target language modeling in MT Son et al. (2012); Kalchbrenner and Blunsom (2013). More recent works use deep learning to model phrase probabilities, e.g. Cho et al. (2014); new reordering models, e.g. Li et al. (2013); or new different features Lu et al. (2014). Different neural architectures to face bilingual translations have been presented in e.g. Sundermeyer et al. (2014); Kalchbrenner and Blunsom (2013).

2.3. Dimesionality reduction techniques for similarity estimation

The field of similarity estimation in continuous space has also advanced in the recent past. The early models based on LSA (Dumais et al., 1988) laid the foundation for dimensionality reduction techniques to incorporate context in form of correlation matrix. Same formulation was exploited by some more advanced linear models such as oriented principle component analysis (OPCA) (Platt et al., 2010) and S2Net (Yih et al., 2011). The other non-linear extensions which outperform to linear counterparts include use of deep autoencoders (Hinton and Salakhutdinov, 2006; Srivastava et al., 2013; Gupta et al., 2014). In this work, we exploit the deep autoencoders based model to estimate source context similarity.

3. Extended Phrase-based Model

This section describes the standard phrase-based SMT system and the methodology of the integration of source contexts in this system both from the theoretical and practical point of view.

3.1. Phrase-based SMT

Given a source string $s_1^J = s_1 \ldots s_j \ldots s_J$ to be translated into a target string $t_1^I = t_1 \ldots t_i \ldots t_I$, a phrase-based SMT system aims to choose, among all possible target strings, the string with the highest probability:

$$\tilde{t}_1^I = \underset{t_1^I}{argmax}\, P(t_1^I | s_1^J)$$

where I and J are the number of words of the target and source sentence, respectively. The phrase-based system segments the source sentence into segments, then translates each segment by using *phrases* which contain source and target sequence of words ($s_1..s_n|||t_1..t_m$). Finally, the system composes the target sentence. Standard implementations of the phrase-based system use several features to give probabilities to combine the relative frequencies together with the: target language model, word and phrase bonus and source-to-target and target-to-source lexical models and reordering model Koehn et al. (2007).

3.2. Theoretical Integration Methodology

The idea of an extended concept of translation unit or phrase (p) is defined by a unit of three elements: *phrase-source* (ps), *phrase-target* (pt) and *source-sentence* (ss).

$$p = \{ps|||pt|||ss\} \quad (1)$$

From this definition identical source-target phrase pairs that have been extracted from different training sentences (or source sentences) are regarded as different translation units. According to this, the relatedness of contexts can be considered as an additional (hereinafter, source-context) feature function (*scf*) for each phrase and input sentence.

$$p = \{ps|||pt|||scf\} \tag{2}$$

The source-context feature function consists of a similarity measurement between the input sentence to be translated and the source context component of the available translation units as illustrated in Fig. 1.

S1: the hotel did not book more rooms
T1: el hotel no reservaba más habitaciones

S2: everybody wants to write a book about himself
T2: todo el mundo quiere escribir un libro sobre sí mismo

Input: i am reading a nice book

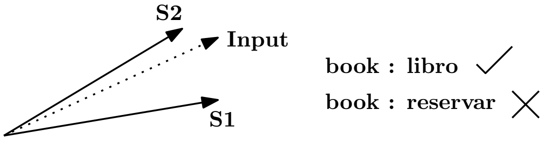

Fig. 1. Illustration of the proposed similarity feature to help choosing translation units.

This *scf* is included for each phrase in addition to the standard feature functions, i.e. conditional (*cp*) and posterior (*pp*) probability, lexical weights (*l*1, *l*2) and phrase bonus (*pb*). Therefore, we are extending the phrases.

$$p = \{ps|||pt|||cp, pp, l1, l2, pb, scf\} \tag{3}$$

This schema is similar to previous work with Banchs and Costa-jussà (2011). Differently from the previous work, for computing similarities between the input sentence to be translated and the original sentences, we compute a cosine distance between the deep representation of sentences, which is explained in Section 4.

3.3. Practical Integration Implementation

The source-context feature function is dynamic because it depends on the input sentence to be translated. At the moment, this feature function is integrated in the standard phrase-based SMT system as described by the following procedure.

Fig. 2 shows the procedure for implementing the source-context feature function. For each training (*ts*) and validation (*vs*) (either development or test) sentence, we compute the similarity measure and build the similarity matrix (*W*) between the training and the validation set. Then, for each sentence in the validation set (vs_n) we extract a phrase list (P_n) that can be used for decoding. Each phrase entry (*p*) in the phrase list is an extended translation unit that contains the training sentence (ts_m) from which it was extracted. Then, the phrase entry is assigned the corresponding source-context similarity from matrix *W* between vs_n and ts_m, which is position W_{mn}. Finally, each sentence in the validation set vs_n is translated with its corresponding extended phrase table (P^*_n) that now includes the source-context feature. The flow of the system is depicted in Fig. 3.

```
$T_M = M$ training sentences ($ts$)
$V_N = N$ validation sentences ($vs$)
for each $ts_m \in T_M$
    for each $vs_n \in V_N$
        $W_{mn} = \omega\ (ts_m, vs_n) =$ similarity$(ts_m, vs_n)$
    end for
end for
for each $vs_n \in V_N$
    $P_n =$ Phrase List $\in T_M$ used for decoding
    $p =$ Phrase Entry $|||\ ts_m \in P_n$
    for each $p \in P_n$
        $p^* \leftarrow p\ |||\ W_{mn}$
    end for
translate $vs_n$ with $P^*_n$
end for
```

Fig. 2. Source-context feature implementation algorithm.

4. Deep Representation of Sentences

We represent the sentences in a latent space through non-linear dimensionality reduction technique. Our method is based on the deep autoencoder architecture which allows to obtain effective low-dimensional embeddings of text data. The autoencoder is a network which tries to learn an approximation of the identity function so as the output is similar to input. The input and output dimensions of the network are the same (n). The autoencoder approximates the identity function in two steps: *i)* reduction, and *ii)* reconstruction. The reduction step takes the input $\mathbf{v} \in \mathbb{R}^n$ and maps it to $\mathbf{h} \in \mathbb{R}^m$ where $m < n$ which can be seen as a function $\mathbf{h} = g(\mathbf{v})$ with $g : \mathbb{R}^n \to \mathbb{R}^m$. On the other hand, the reconstruction step takes the output of the reduction step $\mathbf{h}$ and maps it to $\hat{\mathbf{v}} \in \mathbb{R}^n$ in such a way $\hat{\mathbf{v}} \approx \mathbf{v}$ which is considered as a $\hat{\mathbf{v}} = f(\mathbf{h})$ with function $f : \mathbb{R}^m \to \mathbb{R}^n$. The full autoencoder can be seen as $f(g(\mathbf{v})) \approx \mathbf{v}$.

In a neural network based implementation of the autoencoder, the visible layer corresponds to the input $\mathbf{v}$ and the hidden layer corresponds to $\mathbf{h}$. When the m is sufficently small the autoencoder is able to derive powerful low-dimensional representation of data in the latent space Hinton and Salakhutdinov (2006). There are two variants of autoencoders: *i)* with a single hidden layer, and *ii)* with multiple hidden layers. If there is only one single hidden layer, the optimal solution remains the PCA projection even with the added non-linearities in the hidden layer Bourlard and Kamp (1988). The PCA limitations are overcome by stacking multiple encoders, constituting what is called a deep architecture. This deep construction is what leads

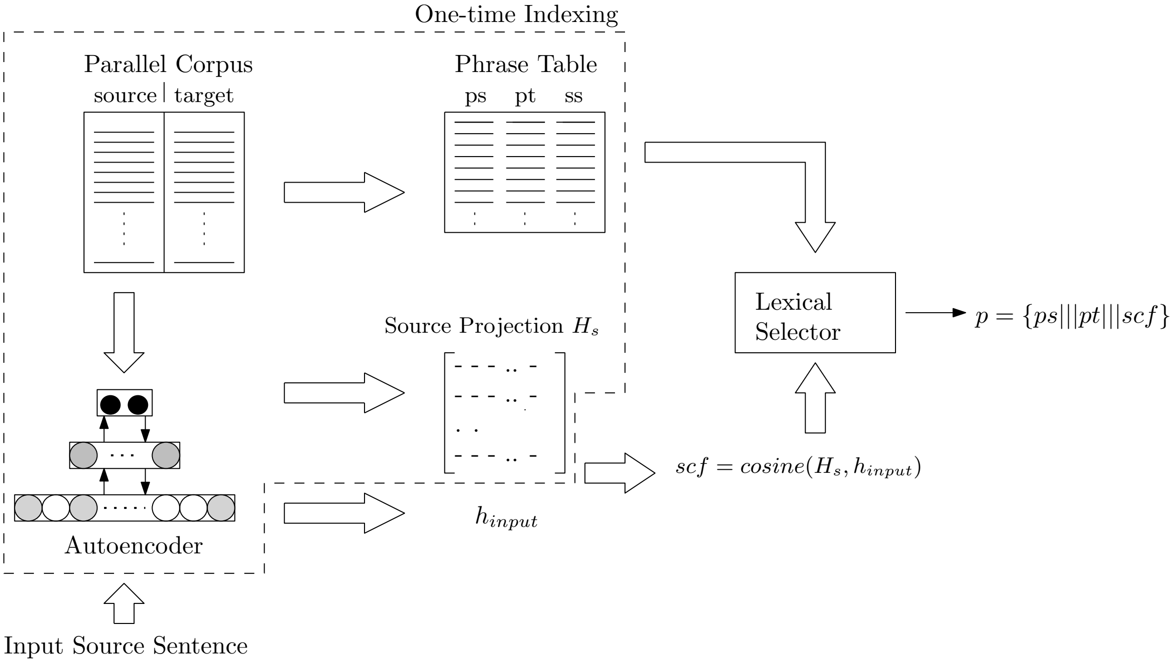

Fig. 3. Workflow of the system.

to a truly non-linear and powerful reduced space representation Hinton and Salakhutdinov (2006). The deep architecture is constituted by stacking multiple restricted boltzmann machines (RBM) on top of each other.

Let visible units $\mathbf{v} \in \{0,1\}^n$ be binary bag-of-words representation of text documents and hidden units $\mathbf{h} \in \{0,1\}^m$ be the hidden latent variables. The energy of the state $\{\mathbf{v},\mathbf{h}\}$ is as follows,

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i,j} v_i h_j w_{ij} \tag{4}$$

where v_i, h_j are the binary states of visible unit i and hidden unit j, a_i, b_j are their biases and w_{ij} is the weight between them.

Then, it becomes easy to sample the data in both directions as shown below,

$$p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j W_{ij}) \tag{5}$$

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i W_{ij}) \tag{6}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid function.

The architecture of the autoencoder is shown in Fig. 4. Latent space representation $\mathbf{h}_s^{(2)}$ for a sentence s can be obtained as shown in Eq. 6. The sentences in the latent space can be compared by means of *cosine* similarity as shown below:

$$\omega(s1, s2) = cosine(\mathbf{h}_{s1}^{(2)}|\mathbf{v}_{s1}, \mathbf{h}_{s2}^{(2)}|\mathbf{v}_{s2}) \tag{7}$$

5. Experiments

This section describes the experimental framework used to test the introduction of deep context features in a standard phrase-based SMT system.

We report details on the data sets used, the baseline system, the training of the deep structure from which similarities among sentences are extracted, the improvements of our technique in terms of BLEU score Papineni et al. (2002) and, finally, the scalability of the technique.

5.1. Data Sets and Baseline

We used an English-to-Spanish parallel corpus extracted from the Bible, which is publicly available and constitutes an excellent corpus for experimenting with and testing the proposed methodology as it provides a rich variety of contexts. The corpus contains around 30,000 sentences of training with around 800,000 words, and 500 sentences each development and test sets. Additionally, as a larger data set, we used an English-to-Hindi corpus available from WMT 2014 Bojar et al. (2014). The training sentences are 300,000 sentences, with 3,500,000 words, 429 sentences of development and 500 sentences of test. Our baseline system is a standard state-of-the-art phrase-based built using Moses toolkit Koehn et al. (2007). We used the following options to train the system, which include: grow-diagonal-final-and word alignment symmetrization, lexicalized reordering, relative frequencies (conditional and posterior probabilities) with phrase discounting, lexical weights and phrase bonus for the translation model (with phrases up to length 10), a 5-gram language model using Kneser-Ney smoothing and a word bonus model. In order to further compare our technique we built a contrastive system with a context

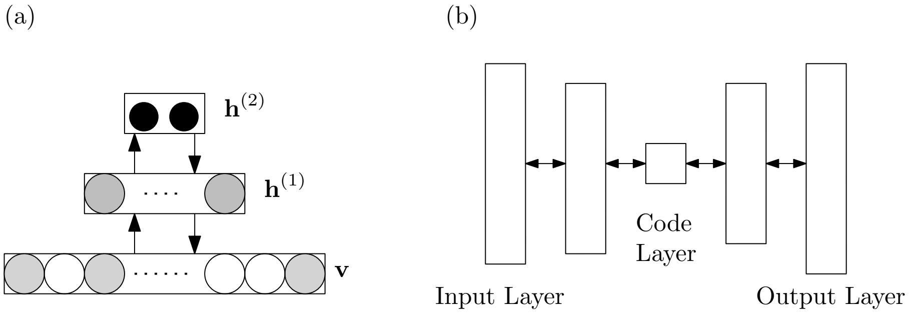

Fig. 4. The architecture of the autoencoder. (a) deep formation of stacked RBMs. (b) Unrolling during the fine-tuning.

feature based on LSA Banchs and Costa-jussà (2011) as another baseline. The systems were computed on a *2Intel Xeon E52670 v3 2,3Ghjz 12N processors server.*

5.2. *Autoencoder training*

To model the sentences in the autoencoder framework we consider the vocabulary after removing the least frequent terms which appear in less than k sentences in the training partition of the dataset. We remove the stopwords and apply stemmer. For Bible and WMT14 dataset, the considered vocabulary sizes (n) are 3543 (k=5) and 7299 (k=20) respectively [1].

The autoencoder was first pretrained using Contrastive Divergence (CD) with step size 1 (Hinton, 2002). Minibatches of size 20 and 100 were used during pretraining and fine-tuning respectively. The architecture of the autoencoder was n-500-128-500-n [2] as shown in Figure 4. Weight decay was used to prevent overfitting. Additionally, in order to encourage sparsity in the hidden units, Kullback-Leibler sparsity regularization was used. We used GPU[3] based implementation of autoencoder to train the models which took around 45 minutes for Bible dataset while around 4.5 hours for WMT14 dataset.

5.3. *Latent Semantic Analysis*

LSA basically performs singular value decomposition of the sentence-term matrix D in the lines of principal component analysis (PCA) (Dumais et al., 1988). LSA obtains top k principal components of D which is considered as projection space and sentences are compared in this space. The inherent idea is semantically similar terms (dimensions of D) will correspond to similar latent components and these sentences are near to each other in the reduced comparison space.

This method can also be looked as eigenproblem which is formulated as below:

$$Cv_j = \lambda_j v_j, \tag{8}$$

where, λ_j is the j^{th} largest eigenvalue, v_j is corresponding eigenvector and C is correlation matrix ($D^T D$). LSA uses top k eigenvectors for projection.

5.4. *Results*

Table 1 shows the improvements in terms of BLEU Papineni et al. (2002) of adding deep context over the baseline system for English-to-Spanish (En2Es) and English-to-Hindi (En2Hi), respectively over development and test sets. Note that the En2Es quality is higher than En2Hi because the former is an easier translation task than the latter and with a higher training corpus. As shown in the Tables, the proposed method performs significantly better than the baseline and than the LSA method for both translation tasks consistently.

	En2Es		**En2Hi**	
	Dev	Test	Dev	Test
baseline	36.81	37.46	9.42	14.99
+LSA	37.20*	37.84*	9.83*	15.12*
+Deep	**37.28***†	**38.19***†	**10.40***†	**15.43***†

Table 1. BLEU scores for En2Es and En2Hi translation tasks. * and † depicts statistical significance (p-value<0.05) *wrt* Baseline and LSA respectively.

It can be noticed that the results from En2Es and En2Hi are consistently improved. We can argue that both Hindi and Spanish have a higher vocabulary variation compared to English, with richer morphology. The benefits of adding source-context information are better reflected in cases where the source phrase can have various target word translations. The improvements in translation proves that the deep representation helps finding the adequate contextual similarities among training and test sentences. BLEU scores show improvement over all tasks and translation directions. Further analysis of the translation outputs presented in Table 2 using ASIYA [4] shows some examples of how the translation is improved in terms of lexical selection which is the goal of the methodology presented in the paper. Examples are shown in Table 2.

In Table 3, we further analyse why our method improves. It can be noticed in the Table 3 that the most probable sense of

[1] The value of k is decided considered from the size of the dataset and the size of vocabulary

[2] Different architectures were tried with a rule of higher layers not larger than the previous layers (because of sparsity in the data) but no statistical difference in results was observed. We also tried three layers n-500-250-128-250-500-n which produced worse results, so we did not go beyond 3-layers.

[3] NVIDIA GeForce GTX Titan with Memory 6 GiB and 2688 CUDA cores

[4] http://www.asiya.lsi.upc.edu

System	Translation
Source	but he brake the bands
Baseline	pero él rompió las tropas
`+Deep`	pero él rompió las **cuerdas**
Reference	pero él rompió las **ataduras**
Source	soft cry from the depth
Baseline	गहराइयों से मूलायम रोने लगते
`+Deep`	गहराइयों से मूलायम **चीख**
Reference	गहराइयों से कोमल **चीख**

Table 2. Manual analysis of translation outputs. Adding the deep feature allows for a more adequate lexical selection.

	cp	*pp*	*scf*
bands|||tropas	0.31	0.17	0.01
bands|||cuerdas	0.06	0.07	0.23
cry|||रोना	0.23	0.06	0.85
cry|||चीख	0.15	0.04	0.90

Table 3. Probability values a phrase-based system) for the word *bands* and two Spanish translations; and the word *cry* a nd two Hindi translations.

bands in our considered dataset is *tropas*, which literally means "troups". The idea of the proposed source-context feature is to use the contextual similarity between the input sentence (IN) and the sentences in the training set as an additional source of information used during decoding. Therefore, given the entire input sentence: *And he was kept bound with chains and in fetters ; and he brake the bands*, the method is be able to infer the correct sense for the word *bands* (i.e. in this case *cuerdas*, which literally means "ropes", a synonym of the reference *ataduras*, which literally means "tying with ropes") by considering its similarity to the training sentences: (S1) *and the lord sent against him bands of the chaldees , and bands of the syrians* and (S2) *they shall put bands upon thee , and shall bind thee with them.* In this case, $\omega(s2, in) > \omega(s1, in)$ as seen in Table 3. Similarly, in the Hindi example, the most frequent sense of word *cry* is रोना, which literally means "to cry" while the example in Table 2 refers to the sense of *cry* as चीख, which means to *scream*. Our method could identify the context and hence the scf(cry|||चीख) > scf(cry|||रोना).

This source-context feature is capable of choosing better translation units given the context but only if the correct translation has been seen in the training data.

5.5. Scalability

There are two components of this method: *i)* Incorporation of source-context features during the tuning phase of MT and projection of training sentences in the latent space; and *ii)* similarity estimation of the input sentence with the training sentences in the latent space. The former step is computationally expensive but being one-time and offline, it is not a big concern. While the similarity estimation is online, it can be very efficiently computed using multi-cores CPU or GPU as it is essentially a matrix multiplication. However, we plan to further integrate this similarity estimation in the translation decoding.

6. Conclusions

This work has shown a novel methodology exploiting deep representation techniques to effectively include a deep learning based contextual similarity estimation method which handles source context and its incorporation in the end-to-end SMT system.

The proposed method shows statistically significant improvements compared to the strong baseline systems in English-to-Spanish and English-to-Hindi translation tasks.

Manual analysis clearly illustrates the advantages in choosing the appropriate translation unit taking into account the information of the input sentence context and the deep relation with the training sentences.

The presented method also scales during the run-time.

Interesting further work would be to include shorter contexts, experiment with deeper auto-encoders and better integrate the dynamic feature into translation decoding. Also, to speed-up search we could divide the feature space in chunks and search hierarchically, perform clustering or use kd-trees.

Acknowledgments

The work has been supported in part by WIQ-EI (IRSES grant n. 269180) and DIANA-APPLICATIONS (TIN2012-38603-C02-01) project and VLC/CAMPUS Microcluster on Multimodal Interaction inIntelligent Systems. The work of the second authors has been supported by Spanish Ministerio de Economía y Competitividad, contract TEC2012-38939-C03-02 as well as from the European Regional Development Fund (ERDF/FEDER) and the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951).

References

Banchs, R.E., Costa-jussà, M.R., 2011. A semantic feature for statistical machine translation, in: Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 126–134.

Bengio, Y., 2012. Deep learning of representations for unsupervised and transfer learning, in: ICML Unsupervised and Transfer Learning, pp. 17–36.

Bojar, O., Diatka, V., Rychl, P., Stranak, P., Suchomel, V., Tamchyna, A., Zeman, D., 2014. Hindencorp - hindi-english and hindi-only corpus for machine translation, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

España Bonet, C., Giménez, J., Màrquez, L., 2009. Discriminative phrase-based models for arabic machine translation. Transactions on Asian Language and Information Processing 8, 15:1–15:20. URL: `http://doi.acm.org/10.1145/1644879.1644882`, doi:`10.1145/1644879.1644882`.

Bourlard, H., Kamp, Y., 1988. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics 59, 291–294. URL: `http://dx.doi.org/10.1007/bf00332918`, doi:`10.1007/bf00332918`.

Carpuat, M., Wu, D., 2005. Word sense disambiguation vs. statistical machine translation, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 387–394. URL: `http://dx.doi.org/10.3115/1219840.1219888`, doi:`10.3115/1219840.1219888`.

Chen, B., Kuhn, R., Foster, G., 2013. Vector space model for adaptation in statistical machine translation, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria. pp. 1285–1293. URL: `http://www.aclweb.org/anthology/P13-1126`.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1724–1734.

Costa-jussà, M.R., Gupta, P., Rosso, P., Banchs, R.E., 2014. English-to-hindi system description for wmt 2014: Deep source-context features for moses, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA. pp. 79–83. URL: http://www.aclweb.org/anthology/W14-3306.

Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R., 1988. Using latent semantic analysis to improve access to textual information, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 281–285. URL: http://doi.acm.org/10.1145/57167.57214, doi:10.1145/57167.57214.

Erhan, D., Bengio, Y., Courville, A.C., Manzagol, P.A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research 11, 625–660.

Glorot, X., Bordes, A., Bengio, Y., 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach, in: ICML, pp. 513–520.

Gupta, P., Bali, K., Banchs, R.E., Choudhury, M., Rosso, P., 2014. Query expansion for mixed-script information retrieval, in: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014, pp. 677–686.

Haque, R., 2010. Integrating Source-Language Context into Log-Linear Models of Statistical Machine Translation. Ph.D. thesis. Dublin City University.

Hasler, E., Blunsom, P., Koehn, P., Haddaw, B., 2014. Dynamic topic adaptation for phrase-based mt, in: Proceedings of the European of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden. URL: http://www.aclweb.org/anthology/P13-1126.

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504 – 507.

Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation 14, 1771–1800.

Kalchbrenner, N., Blunsom, P., 2013. Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA. pp. 1700–1709. URL: http://www.aclweb.org/anthology/D13-1176.

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.

Li, P., Liu, Y., Sun, M., 2013. Recursive autoencoders for ITG-based translation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA. pp. 567–577. URL: http://www.aclweb.org/anthology/D13-1054.

Liu, L., Watanabe, T., Sumita, E., Zhao, T., 2013. Additive neural networks for statistical machine translation, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria. pp. 791–801. URL: http://www.aclweb.org/anthology/P13-1078.

Lu, S., Chen, Z., Xu, B., 2014. Learning new semi-supervised deep autoencoder features for statistical machine translation, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland. pp. 122–132. URL: http://www.aclweb.org/anthology/P14-1012.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318.

Platt, J.C., Toutanova, K., tau Yih, W., 2010. Translingual document representations from discriminative projections, in: EMNLP, pp. 251–261.

Salakhutdinov, R., Hinton, G.E., 2009. Replicated softmax: an undirected topic model, in: NIPS, pp. 1607–1614.

Schwenk, H., Costa-jussà, M.R., Fonollosa, J.A.R., 2006. Continuous space language models for the IWSLT 2006 task, in: 2006 International Workshop on Spoken Language Translation, IWSLT 2006, Keihanna Science City, Kyoto, Japan, November 27-28, 2006, pp. 166–173.

Schwenk, H., R. Costa-jussà, M., R. Fonollosa, J.A., 2007. Smooth bilingual *n*-gram translation, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic. pp. 430–438. URL: http://www.aclweb.org/anthology/D/D07/D07-1045.

Son, L.H., Allauzen, A., Yvon, F., 2012. Continuous space translation models with neural networks, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 39–48. URL: http://dl.acm.org/citation.cfm?id=2382029.2382036.

Srivastava, N., Salakhutdinov, R., Hinton, G.E., 2013. Modeling documents with deep boltzmann machines, in: UAI.

Sundermeyer, M., Alkhouli, T., Wuebker, J., Ney, H., 2014. Translation modeling with bidirectional recurrent neural networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 14–25. URL: http://www.aclweb.org/anthology/D14-1003.

Vaswani, A., Zhao, Y., Fossum, V., Chiang, D., 2013. Decoding with large-scale neural language models improves translation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1387–1392.

Yih, W., Toutanova, K., Platt, J.C., Meek, C., 2011. Learning discriminative projections for text similarity measures, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011, pp. 247–256.

Zamora-Martínez, F., Bleda, M.J.C., Schwenk, H., 2010. N-gram-based machine translation enhanced with neural networks for the french-english btec-iwslt'10 task, in: 2010 International Workshop on Spoken Language Translation, IWSLT 2010, Paris, France, December 2-3, 2010, pp. 45–52. URL: http://www.isca-speech.org/archive/iwslt_10/slta_045.html.