A Semantic Feature for Statistical Machine Translation

Rafael E. Banchs
Institute for Infocomm Research
1 Fusionopolis Way, 21-01, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Marta R. Costa-jussà
Barcelona Media Innovation Centre
Av. Diagonal 177, planta 9, 08018 Barcelona
marta.ruiz@barcelonamedia.org

Abstract

A semantic feature for statistical machine translation, based on Latent Semantic Indexing, is proposed and evaluated. The objective of the proposed feature is to account for the degree of similarity between a given input sentence and each individual sentence in the training dataset. This similarity is computed in a reduced vector-space constructed by means of the Latent Semantic Indexing decomposition. The computed similarity values are used as an additional feature in the log-linear model combination approach to statistical machine translation. In our implementation, the proposed feature is dynamically adjusted for each translation unit in the translation table according to the current input sentence to be translated. This model aims at favoring those translation units that were extracted from training sentences that are semantically related to the current input sentence being translated. Experimental results on a Spanish-to-English translation task on the Bible corpus demonstrate a significant improvement on translation quality with respect to a baseline system.

1 Introduction

In recent years, the statistical approach to machine translation has gained a lot of attention from both the scientific and the commercial perspective. This has basically been a consequence of the increasing availability of bilingual training material as well as the increasing storage and processing capabilities of current computational systems, which have allowed for the construction of machine translation systems with general-public acceptance quality.

For several reasons, the most prominent statistical machine translation paradigm currently used is the phrase-based approach (Koehn *et al.*, 2003), which has been derived from the IBM's word-based approach originally proposed in the early 90's (Brown *et al.*, 1993). This original approach was heavily rooted on the noisy-channel model framework, which, in our view, continues to play an important role in the fundamental conception of current statistical machine translation.

While one of the major assumptions of the noisy-channel model approach is the independence between decoding and source language probabilities, there exists strong evidence on the important role played by source language structure and context within the task of human translation (Padilla & Bajo, 1998). In this sense, the inability of mainstream statistical machine translation to tackle with source-context information in a reliable way has been already recognized as a major drawback of the statistical approach, whereas the use of source-context information has been proven to be effective in the case of example-based machine translation (Carl & Way, 2003). In this regard, attempts for incorporating source-context information into the phrase-based machine translation framework have been already reported (Carpuat & Wu, 2007; Carpuat & Wu, 2008; Haque *et al.*, 2009; España-Bonet *et al.*, 2009; Haque *et al.*, 2010; Costa-jussà & Banchs, 2010). However, as far as we know, no transcendental improvements in performance have been achieved or, at least, reported yet.

In this work, we elaborate deeper on the ideas we have recently presented and discussed in Costa-jussà & Banchs (2010), where we used a similarity metric between the source sentence to be translated and all the sentences in the training set as an addi-

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 126–134,
ACL HLT 2011, Portland, Oregon, USA, June 2011. ©2011 Association for Computational Linguistics

tional feature in the log-linear combination (Och & Ney, 2002) of models of a phrase-based translation system. Such a feature, which is dynamic in the sense that depends on the input sentence to be translated, is intended to favor those translation units which were extracted from training sentences that are similar to the current input sentence over those translation units which were extracted from different or unrelated sentences. Different from our original methodology, where sentence similarities were assessed over a term-document matrix representation for words and statistical classes of words, here we compute sentence similarities in a low-dimensional vector space constructed by means of Latent Semantic Indexing (Landauer *et al.*, 1998).

The rest of the paper is organized as follows. Section 2 presents an overview of some recent approaches attempting to introduce source-context information into the statistical machine translation framework. Then, section 3 introduces the methodology that is proposed and evaluated in this work, and section 4 focuses on some implementation issues. Section 5 describes the experimental settings and results. Section 6 presents a manual evaluation of a selected sample of system translations and discusses the most relevant findings and observations. Finally, section 7 presents the most relevant conclusions of this work and provides guidelines for further research in this area.

2 Related Work

Several attempts for incorporating source-context information into the statistical machine translation framework have been reported in the literature during the last few years. Without attempting to be comprehensive, we provide a brief overlook of some of the most sounded recent works within this area which are relevant to the phrase-based statistical machine translation approach. For a more comprehensive review of the state-of-the-art, the reader can refer to Haque *et al.* (2010).

On the one hand, there are some semantic approaches. In Carpuat & Wu (2007), for instance, word sense disambiguation techniques are introduced into statistical machine translation; and in Carpuat & Wu (2008), dynamically-built context-dependant phrasal translation lexicons are shown to be more useful for phrase-based machine translation than conventional static phrasal translation lexicons, which ignore all contextual information.

On the other hand, there are approaches which use machine learning techniques. In Haque *et al.* (2009), different syntactic and lexical features are proposed for incorporating information about the neighbouring words; and in España-Bonet *et al.* (2009), local classifiers are trained, using linguistic and context information, to translate a phrase.

Finally, our recent approach, which is inspired on information retrieval techniques for measuring the source-context similarity between the input sentence to be translated and the original training material, was presented in Costa-jussà & Banchs (2010). As our present methodology is closely related to this approach, more details are provided in the following section.

3 Proposed Methodology

As already mentioned, the methodology proposed and evaluated in this work is based on the source-context similarity approach we presented in Costa-jussà & Banchs (2010). Different from that work, here we introduce the use Latent Semantic Indexing (Landauer *et al.*, 1998) to construct a vector-space model representation of the data collection in a reduced-dimensionality space before computing source sentence similarities. First, in subsection 3.1, we review the source-context similarity approach. Then, in subsection 3.2 we present the basics of Latent Semantic Indexing.

3.1 The Source-Context Similarity Approach

The method we proposed in Costa-jussà & Banchs (2010) introduces and extended concept of translation unit or phrase by defining a tuple of three elements: phrase-source-side, phrase-target-side, and source-context:

$$TU = \{PSS \,|||\, PTS \,|||\, SC\} \,. \qquad (1)$$

In the most simplistic approach, the source-context element of a given translation unit can be approximated by the complete source sentence the translation unit was originally extracted from. To illustrate this point, consider the following conventional translation unit {*vino*|||*wine*} which has been extracted from the training sentence *sus ojos están brillantes por el vino y sus dientes blancos por la leche* (his eyes shall be red with wine and his teeth white with milk). According to (1), the extended translation unit *TU* is defined as {*vino*|||*wine*|||*sus* *ojos están brillantes por el vino y sus dientes blancos por la leche*}. Notice that, from this definition, identical source-target phrase pairs that have been extracted from different training sentences are regarded as different translation units!

According to this definition, the relatedness of contexts between any translation unit and an input sentence to be translated can be computed by means of some distance or similarity metric over a semantic space representation for sentences. This idea is implemented in practice by means of the following dynamic feature function:

$$F(TU,IN) = SIM(TU,IN) = SIM(SC,IN)\ , \qquad (2)$$

where *TU* refers to a given translation unit, *IN* refers to the input sentence to be translated, *SC* refers to the source-context component of translation unit *TU* (which in our implementation is the source training sentence which the translation unit was extracted from), and *SIM* is a similarity metric over a given model space.

As implied in (2), the source-context feature to be implemented consists of a similarity measurement between the input sentence to be translated *IN* and the source-context component *SC* of the available translation units.

In Costa-jussà & Banchs (2010), we used the cosine of the angle between vectors in a term-sentence matrix representation (Salton *et al.*, 1975) for computing the source-context similarity feature described in (2). In this work, we use Latent Semantic Indexing (Landauer *et al.*, 1998) for projecting the term-sentence matrix representation into a low-dimensional space and use the cosine of the angle between vectors in the resulting reduced space for computing the source-context similarity feature. With this, we expect to reduce the noise resulting from data sparseness problems in the original full-dimensional representation.

To better illustrate the concepts discussed here, let us consider the Spanish word *vino* and the corresponding English translations for its two senses: *wine* and *came*. Both translations can be automatically inferred from training data; and Table 1 illustrates the resulting probability values derived for both senses of the Spanish word *vino* from the actual training dataset used in this work (a detailed description of the dataset is given in section 5).

Notice from the table, how in general the most probable sense of *vino* in our considered dataset is *wine*. This actually happens because the English word *wine* is always related to the Spanish word *vino*, whereas the English word *came* can refer to many different inflections of the same Spanish word: *vine*, *viniste*, *vino*, *vinimos*, *vinieron*, etc.

phrase	$\phi(f\|e)$	$lex(f\|e)$	$\phi(e\|f)$	$lex(e\|f)$
{*vino*|||*wine*}	0.665198	0.721612	0.273551	0.329431
{*vino*|||*came*}	0.253568	0.131398	0.418478	0.446488

Table 1: Actual probability values for the two possible translations of the Spanish word *vino*.

The idea of the proposed source-context feature is to use the contextual similarity between the input sentence to be translated and the sentences in the training dataset as an additional source of information that should be helpful during decoding.

Consider for instance the following two sentences corresponding to the *wine* sense of *vino*:

SC1: No habéis comido pan ni tomado **vino** ni licor , para que sepáis que yo soy Jehovah vuestro Dios . (Ye have not eaten bread , neither have ye drunk **wine** or strong drink : that ye might know that I am the Lord your God .)

SC2: Cuando fue divulgada esta orden , los hijos de Israel dieron muchas primicias de grano , **vino** nuevo , aceite , miel y de todos los frutos de la tierra . (And as soon as the commandment came abroad , the children of Israel brought in abundance the firstfruits of corn , **wine** , and oil , and honey , and of all the increase of the field .)

and the following two sentences corresponding to the *came* sense of *vino*:

SC3: Al tercer día **vino** Jeroboam con todo el pueblo a Roboam , como el rey había hablado diciendo : Volved a mí al tercer día . (So Jeroboam and all the people **came** to Rehoboam the third day , as the king had appointed , saying , Come to me again the third day .)

SC4: Ella **vino** y ha estado desde la mañana hasta ahora . No ha vuelto a casa ni por un momento . (She **came** , and hath continued even from the morning until now , that she tarried a little in the house .)

As the context for a given word is generally determined by its surrounding words, we should be able to infer the correct sense for the word *vino* in a new Spanish sentence by considering its similarity to sentences SC1, SC2, SC3 and SC4. Now, suppose we want to translate the following two input sentences into English:

IN1: Hasta que yo venga y os lleve a una tierra como la vuestra , tierra de grano y de **vino** , tierra de pan y de viñas , tierra de aceite de olivo y de miel . (Until I come and take you away to a land like your own land , a land of corn and **wine** , a land of bread and vineyards , a land of oil olive and of honey .)

IN2: Cuando amanecía , la mujer **vino** y cayó delante de la puerta de la casa de aquel hombre donde estaba su señor , hasta que fue de día . (Then **came** the woman in the dawning of the day , and fell down at the door of the man 's house where her lord was , till it was light .)

We can select the appropriate sense for *vino* in each case by considering the sentence similarity between each of these two sentences and "training" sentences SC1, SC2, SC3 and SC4. The actual similarity values are presented in Table 2.

	SC1	SC2	SC3	SC4
sense	{*vino*|||*wine*}		{*vino*|||*came*}	
IN1	0.0636	**0.2666**	0.0351	0.0310
IN2	0.0023	0.0513	**0.0888**	0.0774

Table 2: Actual similarity values between input and training sentences containing the word *vino*.

As seen from the table, the source-context similarity feature is actually giving preference to the phrase pair {*vino*|||*wine*} in the case of input sentence IN1 and to {*vino*|||*came*} in the case of IN2. Notice that more than one similarity value is generally available for each phrase pair. In our proposed implementation, the largest similarity value is the one that is retained. More details on how we compute these sentence similarities are given in the following subsection.

3.2 Latent Semantic Indexing

Latent Semantic Indexing (Landauer *et al.*, 1998) can be regarded as the text mining equivalent of Principal Component Analysis (Pearson, 1901). Both methods are based on the singular value decomposition (SVD) of a matrix (Golub & Kahan, 1965), according to which a rectangular matrix **X** of dimensions *MxN* can be factorized as follows:

$$\mathbf{X} = \mathbf{U}\, \boldsymbol{\Sigma}\, \mathbf{V}^T , \qquad (3)$$

where **U** and **V** are unitary matrices of dimensions *MxM* and *NxN*, respectively, and **Σ** is a diagonal matrix containing the singular values associated to the decomposition.

According to Landauer *et al.* (1998), a low-dimensional representation of a given document vector ***x*** can be obtained by means of the SVD decomposition depicted in (3) as follows:

$$\boldsymbol{y}^T = \boldsymbol{x}^T\, \mathbf{U}_{\mathbf{MxL}} , \qquad (4)$$

where ***y*** is the *L*-dimensional document vector corresponding to the projection of an *M*-dimensional document vector ***x***, and $\mathbf{U}_{\mathbf{MxL}}$ is a matrix containing the *L* first column vectors of the unitary matrix **U** obtained from (3).

Finally, the feature *F(TU,IN)* described in (2) is implemented as the internal product between normalized versions of the vector projections obtained in (4). In our case, a vector-space model representation is constructed for sentences, instead of documents, and the source-context similarity values between translation units and input sentences are computed accordingly:

$$F\,(TU,\, IN) = \qquad (5)$$
$$< \boldsymbol{sc}^T \mathbf{U}_{\mathbf{MxL}} \,/\, |\boldsymbol{sc}^T \mathbf{U}_{\mathbf{MxL}}| \;,\; \boldsymbol{in}^T \mathbf{U}_{\mathbf{MxL}} \,/\, |\boldsymbol{in}^T \mathbf{U}_{\mathbf{MxL}}| >$$

While the value of *M* is given by the vocabulary size in the data collection under consideration, several implementation questions arise regarding the most appropriate values for *N* (amount of sentences to be used for estimating the projection operator **U**) and *L* (the dimensionality of the reduced space). These and other implementation issues are discussed in detail in the following section.

4 Implementation Issues

This section discusses some important implementation issues that have to be dealt with in order to implement and evaluate the proposed approach. First, in subsection 4.1, the problem of implementing a dynamic feature in a standard phrase-based machine translation framework is discussed. Then, in subsections 4.2 and 4.3, the problems of determining the amount of data required for estimating the Latent Semantic Indexing projection operator and the most appropriate dimensionality size for the reduced space representation are discussed.

4.1 Implementing a Dynamic Feature

As defined in (2), the value of the proposed source-context similarity feature depends on each individual input sentence to be translated by the system. This definition implies a major difference between this feature and other conventional phrase-based translation features: it is a dynamic feature in the sense that it cannot be computed in advance before the input sentences to be translated are known.

This on-the-fly requirement, along with the extended translation unit definition presented in (1), makes it not possible to directly implement the proposed methodology within a standard phrase-based machine translation framework such as MOSES (Koehn *et al.*, 2007). As it is not our intention to develop a customized decoding tool for implementing and testing our proposed feature, we followed or previous implementation of an off-line version of the proposed methodology (Costa-jussà & Banchs, 2010), which, although very inefficient in the practice, allows us to evaluate the impact of the source-context feature on a state-of-the-art phrase-based translation system.

According to this, our practical implementation is a follows:

- Two sentence similarity matrices are computed: one between sentences in the development and training sets, and the other between sentences in the test and training datasets.
- Each matrix entry m_{ij} should contain the similarity score between the i^{th} sentence in the training set and the j^{th} sentence in the development (or test) set.
- For each sentence s in the test and development sets, a phrase list L_S of all potential phrases that can be used during decoding is extracted from the aligned training set.
- The corresponding source-context similarity values are assigned to each phrase in lists L_S according to values in the corresponding similarity matrices.
- Each phrase list L_S is collapsed into a phrase table T_S by removing repetitions (when removing repeated entries in the list, the largest value of the source-context similarity feature is retained).
- Each phrase table is completed by adding standard feature values (which are computed in the standard manner).
- MOSES is used on a sentence-per-sentence basis, using a different translation table for each development (or test) sentence.

4.2 Dataset for Latent Semantic Indexing

Another important implementation issue that requires attention is the computation of the Singular Value Decomposition described in (3). Ideally, the term-sentence matrix **X** to be decomposed should include all available data, i.e. training, development and test sentences; however, in the practice, this is not possible because of two reasons. First, the sizes of typical datasets and vocabularies used in statistical machine translation systems are large enough to make Singular Value Decomposition unfeasible from a computational point of view[1]. Second, in a practical application system, the "test set" is actually unknown during the system construction and training phases. In this way, a realistic implementation should be able to work with previously unseen data.

In order to overcome the problem of applying the Singular Value Decomposition described in (3) to the full term-sentence matrix of all available data, we implemented an approximated procedure. In our approximation, we compute the similarity matrix between two set of sentences as the average of several similarity matrices that are computed over reduced space projections estimated with different random samples of the training data sentences. In this way, our source-context similarity feature, previously defined in (5), becomes:

$$F\ (TU, IN) \approx \qquad (6)$$
$$1/K\ \Sigma_{\mathrm{k}} < \boldsymbol{sc}^{\mathrm{T}}\mathbf{U}^{\mathbf{k}}_{\mathbf{MxL}} / |\boldsymbol{sc}^{\mathrm{T}}\mathbf{U}^{\mathbf{k}}_{\mathbf{MxL}}|\ ,\ \boldsymbol{in}^{\mathrm{T}}\mathbf{U}^{\mathbf{k}}_{\mathbf{MxL}} / |\boldsymbol{in}^{\mathrm{T}}\mathbf{U}^{\mathbf{k}}_{\mathbf{MxL}}| >$$

where $\mathbf{U}^{\mathbf{k}}_{\mathbf{MxL}}$ refers to a projection operator that has been computed by means of the Singular Valued Decomposition of a term-sentence matrix $\mathbf{X}^{\mathbf{k}}$ constructed with a random sample of N sentences. Note that a total of K different similarity scores are averaged in (6).

In order to evaluate the variability of the similarity values estimated by this approximation, several experiments were conducted for different values of N and L, where the variance of the estimates over $K=10$ different realizations were computed. Figure 1 shows the resulting standard deviations for similarity values estimated for different values of L when varying N (upper panel), and for different values of N when varying L (lower panel).

As seen from the figure, the range $500<N<1000$ seems to constitute a good compromise between the size of selected random sentence sets and the observed variability for similarity value estimates, as it provides a significant reduction in the computed standard deviations with respect to $N=100$, and not important improvement is observed when

[1] Even in the case of a small dataset such as the one considered here (see details in section 5) the Singular Value Decomposition of the full term-sentence matrix can take several weeks to be completed in and standard Linux-based server.

N>1000. According to this, we selected *N=1000* for our proposed approximation described in (6).

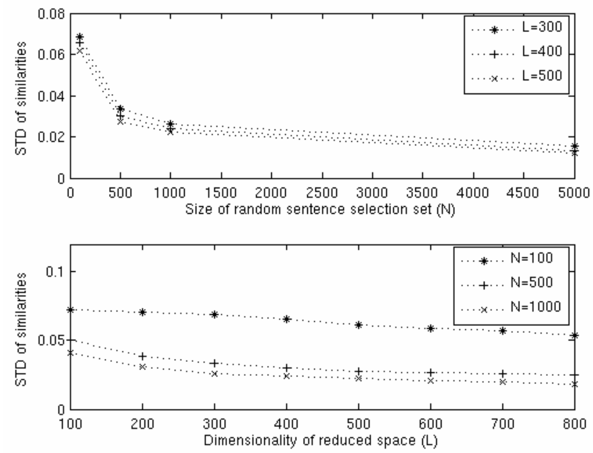

Figure 1: Standard deviations (STD) for similarity values between development and test datasets (described in section 5) estimated for different values of L when varying N (upper panel), and for different values of N when varying L (lower panel). In all cases K=10.

4.3 Reduced Space Dimensionality

The third and final implementation issue to be discussed is the selection of the reduced space dimensionality. It have been reported in the literature that dimensionality reduction, by means of Latent Semantic Indexing, into the range between *100* and *1000* provides good space representations for word and sentence association applications (Landauer *et al.*, 1998). Although it is reasonable to assume this condition to be valid also for the application under consideration, we conducted a more detailed exploratory analysis for selecting the dimensionality *L* to be used in our experiments.

First, we studied the distributions of context-similarity values computed according to (6) over the available data. Figure 2 shows the average distributions of similarities between sentences in the development and training datasets (see data description in section 5) at different dimensionality values. As can be seen from the figure, a dimensionality value of *L=100* exhibits a very nice distribution of similarity values; however, according to the results depicted in Figure 1 (lower panel), the variability of estimates for such a low dimensionality is relatively high. On the other hand, notice again from Figure 2, how a much larger dimensionality value such as *L=5000* already starts to exhibit a distribution of similarities that is heavily biased towards the low similarity region. According to this result, and taking also into account the results in Figure 1, we finally decided setting the dimensionality of the reduced space to *L=500.*

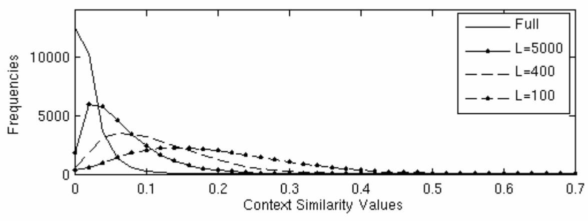

Figure 2: Average distributions of similarity values between development and training sentences computed at different dimensionality values. For all cases presented here N=500 and K=10.

5 Experimental Work

This section describes the experimental work conducted to evaluate the incidence of the proposed source-context similarity feature on translation quality for a state-of-the-art phrase-based statistical machine translation. First, subsection 5.1 describes the dataset and experimental setting. Then, subsection 5.2 presents and discusses the results.

5.1 Experimental Setting

The proposed methodology is evaluated on the Bible dataset (Chew *et al.*, 2006) Spanish-to-English translation task, using the MOSES framework as baseline phrase-based statistical machine translation system (Koehn *et al.*, 2007). Table 3 presents the main statistics of the bilingual corpus used.

dataset	lang.	sentences	tokens	vocab	av. lenght
Train	Spa	28,887	781,113	28,178	27
Train	Eng	28,887	848,776	13,126	29
Test	Spa	500	13,312	2,879	27
Test	Eng	500	14,562	2,156	29
Dev	Spa	500	13,170	2,862	26
Dev	Eng	500	14,537	2,095	29

Table 3: Main statistics of the bilingual corpus under consideration (number of sentences, tokens, vocabulary, and average sentence length)

Regarding the baseline system, we used the default parameters of MOSES, which include the grow-final-diagonal alignment symmetrisation, the lexicalized reordering, a 5-gram language model using Kneser-Ney smoothing, and phrases up to length 10, among others. The optimization was done using the standard MERT procedure (Och & Ney, 2002).

5.2 Experimental Results

Table 4 presents the translation BLEU, measured over the development and test sets, for three different system implementations: the baseline system, a second system implementing the source-context similarity feature over the full-dimensional vector space (FVS), just as we implemented it in Costa-jussà & Banchs (2010), and a third system implementing the source-context similarity feature based on Latent Semantic Indexing (LSI).

	Development	Test
Baseline	39.92	38.92
Source-context (FVS)	40.61	39.43
Source-context (LSI)	**40.80**	**39.86**

Table 4: BLEU scores over development and test datasets corresponding to three system implementations: baseline, and source-context similarity feature at full-dimensional vector space (FVS) and by means of Latent Semantic Indexing (LSI).

As seen from the table, the system implementing the Latent Semantic Indexing based source-context similarity feature outperforms the baseline system by almost one absolute BLEU point, and the full-dimensional vector space system by some less than a half absolute BLEU point. An analysis of significance (Koehn, 2004) showed that the differences among the systems are statistically significant.

A more comprehensive manual analysis of both the baseline and source-context LSI system outputs was required to better asses the incidence of the implemented source-context similarity feature on the generated translations. The result of this analysis is presented in the following section.

6 Manual Evaluation

This section presents and discusses the results of a manual evaluation that was conducted over a sample set of translations. Previous to the manual evaluation, we performed a sentence-based automatic evaluation using BLEU for the *500* sentences in the test dataset. We obtained that our proposed approach is better than the baseline system in *208* sentences, while the baseline is better than our system in *173* sentences and the remaining *119* had the same BLEU scores.

Some output sentences were randomly selected, regardless of which system performed better, for conducting a manual inspection. From these sentences, we have extracted some segments that illustrate specific cases in which our proposed source-context feature is actually helping to select a better translation unit according to the context of the input sentence being translated. Five of these segments are presented in Table 5, where the relevant fragments within the segments are shown in bold.

Example 1	
source	No des sueño a **tus ojos** ni dejes dormitar tus párpados .
reference	Give not sleep to **thine eyes** , nor slumber to thine eyelids .
baseline	Not sleep in **thy sight** , Let neither slumber thy eyelids .
LSI-context	Give not sleep to **thine eyes** neither slumber , Let thine eyelids .
Example 2	
source	Entonces ellos se acercaron , **echaron mano a** Jesús y le prendieron …
reference	Then came they , and **laid hands on** Jesus , and took him …
baseline	And they came near , and **cast hand to** Jesus , and took him …
LSI-context	And they came near , and **laid hands on** Jesus , and took him …
Example 3	
source	Y al tercer día , he aquí que un hombre **vino** del campamento de Saúl …
reference	It came even to pass on the third day , that , behold , a man **came** out of the camp from Saul …
baseline	And the third day , behold , a man **wine** of the camp of Saul …
LSI-context	And the third day , behold , there **came** a man of the camp of Saul …
Example 4	
source	… **sed** confortados ; **sed** de un mismo sentir …
reference	… **be** of good comfort , **be** of one mind …
baseline	… **thirst** confortados ; **thirst** of one mind 's sake …
LSI-context	… **be** ye confortados : **be** ye of one mind 's sake …
Example 5	
source	… según sus familias , según sus idiomas , en **sus territorios** y en sus naciones .
reference	… after their families , after their tongues , in **their countries** , and in their nations .
baseline	… according to their families , after their tongues , in **their coasts** , and in their nations .
LSI-context	… after their families , after their tongues , in **their lands** , and in their nations .

Table 5: Sample segments where the LSI-based source-context feature has helped to accomplish better translation unit selections.

As seen from the table, the LSI-based source-context system is clearly accomplishing more appropriate unit selections. However, in most of the cases this does not imply either a better overall translation or a closer match to the available reference translation. This can explain the relative low BLEU gain achieved by the method.

Similarly, we also extracted some segments that illustrate specific cases in which our proposed source-context feature fails in helping to select a better translation unit. Table 6 presents four of these cases.

Example 1	
source	… yo he sido enviado con **malas noticias** para ti .
reference	… for I am sent to thee with **heavy tidings** .
baseline	… for I have sent with **evil tidings** unto thee .
LSI-context	… I am sent with **evil tidings** unto thee .
Example 2	
source	… heredad de Jehovah son los hijos ; recompensa es el **fruto del vientre** .
reference	… children are an heritage of the Lord : and the **fruit of the womb** is his reward .
baseline	… the inheritance of the Lord , are the children ; reward is the **fruit of the belly** .
LSI-context	… the inheritance of the Lord are the children , and reward is the **fruit of the belly** .
Example 3	
source	… y que **había enaltecido su reino** por amor a su pueblo Israel .
reference	… and that **he had exalted his kingdom** for his people Israel 's sake .
baseline	… and for **his kingdom was lifted up** his people Israel .
LSI-context	… and for **his kingdom was lifted up** unto his people Israel .
Example 4	
source	Y sucederá que a causa de la **abundancia de leche** , comerá leche cuajada …
reference	And it shall come to pass , for the **abundance of milk** that he shall eat butter …
baseline	And it shall come to pass , that by reason of the **multitude of milk** , shall eat with milk cuajada …
LSI-context	And it shall come to pass by reason of the **multitude of milk** , and shall eat with milk cuajada …

Table 6: Sample segments where the LSI-based source-context feature has failed to accomplish better translation unit selections.

In the latter examples in Table 6, the proposed source-context feature is clearly failing to provide better lexical selections. In some cases, this seems to be due to the lack of enough source-context information in the input sentence to be translated. However, in other cases, it is because the source-context feature alone is not able to compensate the system's bias towards more frequent translations.

7 Conclusions and Future Work

A new semantically-motivated feature for statistical machine translation based on Latent Semantic Indexing has been proposed and evaluated. The objective of the proposed feature is to account for the degree of similarity between a given input sentence and each individual sentence in the training dataset. This similarity is computed in a reduced vector-space constructed by means of the Latent Semantic Indexing decomposition.

The computed similarity values are used as an additional feature in the log-linear model combination approach to statistical machine translation. In our implementation, the proposed feature is dynamically adjusted for each translation unit in the translation table according to the current input sentence to be translated.

Experimental results on a Spanish-to-English translation task on the Bible corpus showed significant improvements of almost *1* and *0.5* absolute BLEU points with respect to a baseline system and a similar system evaluating sentence similarity at the full-dimensional vector space, respectively. A manual evaluation revealed that the proposed feature is actually helping the translation system to perform a better selection of translation units on a semantic basis.

As future work, we intend to evaluate different association and distance metrics, as well as to extend the current notion of source-context from the input sentence to be translated to any other kind of available information beyond the input sentence limits. Similarly, different paradigms of semantic space representations, including those statistically motivated, will be studied and evaluated.

Implementation issues are also to be revisited for better evaluating the impact of both the amount of training data and the dimensionality of the reduced space on the method's performance. Finally, an on-line version of the method must be implemented in order to be able to evaluate the proposed methodology over larger data collections.

Acknowledgments

The authors would like to thank the Institute for Infocomm Research, as well as Barcelona Media Innovation Centre and the Juan de la Cierva fellowship program, for their support and permission to publish this work.

References

Brown, P., Della-Pietra, S., Della-Pietra, V., Mercer, R. (1993) The Mathematics of Statistical Machine Translation: Computational Linguistics 19(2), 263--311

Carl, M., Way, A. (2003) Recent Advances in Example-Based Machine Translation. Kluwer Academic

Carpuat, M., Wu, D. (2007) How Phrase Sense Disambiguation Outperforms Word Sense Disambiguation for Statistical Machine Translation. In: 11th International Conference on Theoretical and Methodological Issues in Machine Translation. Skovde

Carpuat, M., Wu, D. (2008) Evaluation of Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation. In: 6th International Conference on Language Resources and Evaluation (LREC). Marrakech

Chew, P. A., Verzi, S. J., Bauer, T. L., McClain, J. T. (2006) Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. In: Workshop on Multilingual Language Resources and Interoperability, pp. 68--74, Sydney

Costa-jussà, M. R., Banchs, R.E. (2010) A Vector-Space Dynamic Feature for Phrase-Based Statistical Machine Translation. Journal of Intelligent Information Systems

España-Bonet, C., Gimenez, J., Marquez, L. (2009) Discriminative Phrase-Based Models for Arabic Machine Translation. ACM Transactions on Asian Language Information Processing Journal (Special Issue on Arabic Natural Language Processing)

Golub, G. H., Kahan, W. (1965) Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics: Numerical Analysis 2(2), 205--224

Haque, R., Naskar, S. K., Ma, Y., Way, A. (2009) Using Supertags as Source Language Context in SMT. In: 13th Annual Conference of the European Association for Machine Translation, pp. 234--241. Barcelona

Haque, R., Naskar, S. K., van den Bosh, A., Way, A. (2010) Supertags as Source Language Context in Hierarchical Phrase-Based SMT. In: 9th Conference of the Association for Machine Translation in the Americas (AMTA)

Koehn, P., Och, F. J., Marcu, D. (2003) Statistical Phrase-Based Translation. In: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP), pp. 48--54. Edmonton

Koehn, P. (2004) Statistical Significance Test for Machine Translation Evaluation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP)

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: 45th Annual Metting of the Association for Computational Linguistics, pp. 177--180. Prague

Landauer, T. K., Laham, D., Foltz, P. (1998) Learning Human-Like Knowledge by Singular Value Decomposition: A Progress Report. In: Conference on Advances in Neural Information Processing Systems, pp. 45--51. Denver

Landauer, T. K., Foltz, P.W., Laham, D. (1998) Introduction to Latent Semantic Analysis. Discourse Processes 25, 259--284

Och, F. J., Ney, H. (2002) Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 295--302

Padilla, P., Bajo, T. (1998) Hacia un Modelo de Memoria y Atención en la Interpretación Simultánea. Quaderns: Revista de Traducció 2, 107--117

Pearson, K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 2(6), 559--572

Salton, G., Wong, A., Yang, C. S. (1975) A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613--620