Manifold Learning for the Semi-Supervised Induction of FrameNet Predicates: An Empirical Investigation

Danilo Croce and **Daniele Previtali**
{croce,previtali}@info.uniroma2.it
Department of Computer Science, Systems and Production
University of Roma, *Tor Vergata*

Abstract

This work focuses on the empirical investigation of distributional models for the automatic acquisition of frame inspired predicate words. While several semantic spaces, both word-based and syntax-based, are employed, the impact of geometric representation based on dimensionality reduction techniques is investigated. Data statistics are accordingly studied along two orthogonal perspectives: Latent Semantic Analysis exploits global properties while Locality Preserving Projection emphasizes the role of local regularities. This latter is employed by embedding prior FrameNet-derived knowledge in the corresponding non-euclidean transformation. The empirical investigation here reported sheds some light on the role played by these spaces as complex kernels for supervised (i.e. Support Vector Machine) algorithms: their use configures, as a novel way to semi-supervised lexical learning, a highly appealing research direction for knowledge rich scenarios like FrameNet-based semantic parsing.

1 Introduction

Automatic Semantic Role Labeling (SRL) is a natural language processing (NLP) technique that maps sentences to semantic representations and identifies the semantic roles conveyed by sentential constituents (Gildea and Jurafsky, 2002). Several NLP applications have exploited this kind of semantic representation ranging from Information Extraction (Surdeanu et al., 2003; Moschitti et al., 2003)) to Question Answering (Shen and Lapata, 2007), Paraphrase Identification (Pado and Erk, 2005), and the modeling of Textual Entailment relations (Tatu and Moldovan, 2005). Large scale annotated resources have been used by Semantic Role Labeling methods: they are commonly developed using a supervised learning paradigm where a classifier learns to predict role labels based on features extracted from annotated training data. One prominent resource has been developed under the Berkeley FrameNet project as a semantic lexicon for the core vocabulary of English, according to the so-called $frame$ semantic model (Fillmore, 1985). Here, a frame is a conceptual structure modeling a prototypical situation, evoked in texts through the occurrence of its lexical units (LU) that linguistically expresses the situation of the frame. Lexical units of the same frame share semantic arguments. For example, the frame KILLING has lexical units such as *assassin, assassinate, blood-bath, fatal, murderer, kill* or *suicide* that share semantic arguments such as KILLER, INSTRUMENT, CAUSE, VICTIM. The current FrameNet release contains about 700 frames and 10,000 LUs. A corpus of 150,000 annotated examples sentences, from the British National Corpus (BNC), is also part of FrameNet.

Despite the size of this resource, it is under development and hence incomplete: several frames are not represented by evoking words and the number of annotated sentences is unbalanced across frames. It is one of the main reason for the performance drop of supervised SRL systems in out-of-domain scenarios (Baker et al., 2007) (Johansson and Nugues, 2008). The limited coverage of FrameNet corpus is even more noticeable for the LUs dictionary: it only contains 10,000 lexical units, far less than the 210,000 entries in WordNet 3.0. For example, the lexical unit *crown*, according to the annotations, evokes the ACCOUTREMENT frame. It refers to a particular sense: according to WordNet, it is "an ornamental jeweled headdress signifying sovereignty". According to the same lexical resource, this LU has 12 lexical senses and the first one (i.e. "The Crown

Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, ACL 2010, pages 7–16,
Uppsala, Sweden, 16 July 2010.

(or the reigning monarch) as the symbol of the power and authority of a monarchy") could evoke other frames, like LEADERSHIP. In (Pennacchiotti et al., 2008) and (De Cao et al., 2008), the problem of LU automatic induction has been treated in a semi-supervised fashion. First, LUs are modeled by exploiting the distributional analysis of an unannotated corpus and the lexical information of WordNet. These representations were used in order to find out frames potentially evoked by novel words in order to extend the FrameNet dictionary limiting the effort of manual annotations.

In this work the distributional model of LUs is further developed. As in (Pennacchiotti et al., 2008), several word spaces (Pado and Lapata, 2007) are investigated in order to find the most suitable representation of the properties which characterize a frame. Two dimensionality reduction techniques are applied here in this context. *Latent Semantic Analysis* (Landauer and Dumais, 1997) uses the Singular Value Decomposition to find the best subspace approximation of the original word space, in the sense of minimizing the global reconstruction error projecting data along the directions of maximal variance. *Locality Preserving Projection* (He and Niyogi, 2003) is a linear approximation of the nonlinear Laplacian Eigenmap algorithm: its locality preserving properties allows to add a set of constraints forcing LUs that belong to the same frame to be near in the resulting space after the transformation. LSA performs a global analysis of a corpus capturing relations between LUs and removing the noise introduced by spurious directions. However it risks to ignore lexical senses poorly represented into the corpus. In (De Cao et al., 2008) external knowledge about LUs is provided by their lexical senses from a lexical resource (e.g WordNet). In this work, prior knowledge about the target problem is directly embedded into the space through the LPP transformation, by exploiting locality constraints. Then a Support Vector Machine is employed to provide a robust acquisition of lexical units combining global information provided by LSA and the local information provided by LPP into a complex kernel function.

In Section 2 related work is presented. In Sections 3 the investigated distributional model of LUs is presented as well as the dimensionality reduction techniques. Then, in Section 4 the experimental investigation and comparative evaluations are reported. Finally, in Section 5 we draw final conclusions and outline future work.

2 Related Work

As defined in (Pennacchiotti et al., 2008), LU induction is the task of assigning a generic lexical unit not yet present in the FrameNet database (the so-called *unknown LU*) to the correct frame(s). The number of possible classes (i.e. frames) and the multiple assignment problem make it a challenging task. LU induction has been integrated at SemEval-2007 as part of the Frame Semantic Structure Extraction shared task (Baker et al., 2007), where systems are requested to assign the correct frame to a given LU, even when the LU is not yet present in FrameNet. Several approaches show low coverage (Johansson and Nugues, 2007) or low accuracy, like (Burchardt et al., 2005). This task is presented in (Pennacchiotti et al., 2008) and (De Cao et al., 2008), where two different models which combine distributional and paradigmatic (i.e. lexical) information have been discussed. The distributional model is used to select a list of frame suggested by the corpus' evidences and then the plausible lexical senses of the unknown LU are used to re-rank proposed frames.

In order to exploit prior information provided by the frame theory, the idea underlying is that semantic knowledge can be embedded from external sources (i.e the FrameNet database) into the distributional model of unannotated corpora. In (Basu et al., 2006) a limited prior knowledge is exploited in several clustering tasks, in term of pairwise constraints (i.e., pairs of instances labeled as belonging to same or different clusters). Several existing algorithms enhance clustering quality by applying supervision in the form of constraints. These algorithms typically utilize the pairwise constraints to either modify the clustering objective function or to learn the clustering distortion measure. The approach discussed in (Basu et al., 2006) employs Hidden Markov Random Fields (HMRFs) as a probabilistic generative model for semi-supervised clustering, providing a principled framework for incorporating constraint-based supervision into prototype-based clustering.

Another possible approach is to directly embed the prior-knowledge into data representations. The main idea is to employ effective and efficient algorithms for constructing nonlinear low-dimensional manifolds from sample data points embedded

in high-dimensional spaces. Several algorithms are defined, including Isometric feature mapping (ISOMAP) (Tenenbaum et al., 2000), Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Local Tangent Space alignment (LTSA) (Zhang and Zha, 2004) and Locality Preserving Projection (LPP) (He and Niyogi, 2003) and they have been successfully applied in several computer vision and pattern recognition problems. In (Yang et al., 2006) it is demonstrated that basic nonlinear dimensionality reduction algorithms, such as LLE, ISOMAP, and LTSA, can be modified by taking into account prior information on exact mapping of certain data points. The sensitivity analysis of these algorithms shows that prior information improves stability of the solution. In (Goldberg and Elhadad, 2009), a strategy to incorporate lexical features into classification models is proposed. Another possible approach is the strategy pursued in recent works on deep learning techniques to NLP tasks. In (Collobert and Weston, 2008) a unified architecture for NLP that learns features relevant to the tasks at hand given very limited prior knowledge is presented. It embodies the idea that a multitask learning architecture coupled with semi-supervised learning can be effectively applied even to complex linguistic tasks such as Semantic Role Labeling. In particular, (Collobert and Weston, 2008) proposes an embedding of lexical information using Wikipedia as source, and exploits the resulting language model for the multitask learning process. The extensive use of unlabeled texts allows to achieve a significant level of lexical generalization in order to better capitalize on the smaller annotated data sets.

3 Geometrical Embeddings as models of Frame Semantics

The aim of this distributional approach is to model frames in semantic spaces where words are represented from the distributional analysis of their co-occurrences over a corpus. Semantic spaces are widely used in NLP for representing the meaning of words or other lexical entities. They have been successfully applied in several tasks, such as information retrieval (Salton et al., 1975) and harvesting thesauri (Lin, 1998). The fundamental intuition is that the meaning of a word can be described by the set of textual contexts in which it appears (*Distributional Hypothesis* as described in (Harris, 1964)), and that words with similar vectors are semantically related. Contexts are words appearing together with a LU: such a space models a generic notion of semantic relatedness, i.e. two LUs spatially close in the space are likely to be either in paradigmatic or syntagmatic relation as in (Sahlgren, 2006). Here, LUs delimit subspaces modeling the prototypical semantic of the corresponding evoked frames and novel LUs can be induced by exploiting their projections.

Since a semantic space supports the language in use from the corpus statistics in an unsupervised fashion, vectors representing LUs can be characterized by different distributions. For example, LUs of the frame KILLING, such as *bloodbath*, *crucify* or *fratricide*, are statistically inferior in a corpus if compared to a wide-spanning term as *kill*. Moreover other ambiguous LUs, as *liquidate* or *terminate*, could appear in sentences evoking different frames. These problems of data-sparseness and distribution noise can be overcome by applying space transformation techniques augmenting the space expressiveness in modeling frame semantics. Semantic space models very elegantly map words in vector spaces (there are as many dimensions as words in the dictionary) and LUs collections into distributions of datapoints. Every distribution implicitly expresses two orthogonal facets: global properties, as the occurrence scores computed for terms across the entire collection (irrespectively from their word senses or evoking situation) and local regularities, for example the existence of subsets of terms that tend to be used every time a frame manifests. These also tend to be closer in the space and should be closer in the transformed space too. Another important aspect that a transformation could account is external semantic information. In the new space, prior knowledge can be exploited to gather a more regular LUs representation and a clearer separation between subspaces representing different frame semantics.

In the following sections the investigated distributional model of LUs will be discussed. As many criteria can be adopted to define a LU context, one of the goals of this investigation is to find a co-occurrence model that better captures the notion of frames, as described in Section 3.1. Then, two dimensionality reduction techniques, exploiting semantic space distributions to improve frames representation, are discussed. In Section 3.2 the role of global properties of data statistics will be

investigated through the Latent Semantic Analysis while in Section 3.3 the Locality Preserving Projection algorithm will be discussed in order to combine prior knowledge about frames with local regularities of LUs obtained from text.

3.1 Choosing the space

Different types of context define spaces with different semantic properties. Such spaces model a generic notion of *semantic relatedness*. Two LUs close in the space are likely to be related by some type of generic semantic relation, either paradigmatic (e.g. synonymy, hyperonymy, antonymy) or syntagmatic (e.g. meronymy, conceptual and phrasal association), as observed in (Sahlgren, 2006). The target of this work is the construction of a space able to capture the properties which characterize a frame, assuming those LUs in the same frame tend to be either co-occurring or substitutional words (e.g. *murder/kill*). Two traditional word-based co-occurrence models capture the above property:

Word-based space: Contexts are words, as lemmas, appearing in a n-window of the LU. The window width n is a parameter that allows the space to capture different aspects of a frame: higher values risk to introduce noise, since a frame could not cover an entire sentence, while lower values lead to sparse representations.

Syntax-based space: Contexts words are enriched through information about syntactic relations (e.g. *X-VSubj-killer* where X is the LU), as described in (Pado and Lapata, 2007). Two LUs close in the space are likely to be in a paradigmatic relation, i.e. to be close in an IS-A hierarchy (Budanitsky and Hirst, 2006; Lin, 1998). Indeed, as contexts are syntactic relations, targets with the same part of speech are much closer than targets of different types.

3.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an algorithm presented in (Furnas et al., 1988) afterwards diffused by Landauer (Landauer and Dumais, 1997): it can be seen as a variant of the Principal Component Analysis idea. LSA aims to find the best subspace approximation to the original word space, in the sense of minimizing the global reconstruction error projecting data along the directions of maximal variance. It captures term (semantic) dependencies by applying a matrix decomposition process called Singular Value Decomposition (SVD). The original term-by-term matrix M is transformed into the product of three new matrices: U, S, and V so that $M = USV^T$. Matrix M is approximated by $M_l = U_l S_l V_l^T$ in which only the first l columns of U and V are used, and only the first l greatest singular values are considered. This approximation supplies a way to project term vectors into the l-dimensional space using $Y_{terms} = U_l S_l^{1/2}$. Notice that the SVD process accounts for the eigenvectors of the entire original distribution (matrix M). LSA is thus an example of a decomposition process strongly dependent on a global property. The original statistical information about M is captured by the new l-dimensional space which preserves the global structure while removing low-variant dimensions, i.e. distribution noise. These newly derived features may be thought of as artificial concepts, each one representing an emerging meaning component as a linear combination of many different words (i.e. contexts). Such contextual usages can be used instead of the words to represent texts. This technique has two main advantages. First, the overall computational cost of the model is reduced, as similarities are computed on a space with much fewer dimensions. Secondly, it allows to capture second-order relations among LUs, thus improving the quality of the similarity measure.

3.3 The Locality Preserving Projection Method

An alternative to LSA, much tighter to local properties of data, is the Locality Preserving Projection (LPP), a linear approximation of the non-linear Laplacian Eigenmap algorithm introduced in (He and Niyogi, 2003). LPP is a linear dimensionality reduction method whose goal is, given a set of LUs $x_1, x_2, .., x_m$ in R^n, to find a transformation matrix A that maps these m points into a set of points $y_1, y_2, .., y_m$ in R^k ($k \ll n$). LPP achieves this result through a cascade of processing steps described hereafter.

Construction of an Adjacency graph. Let G denote a graph with m nodes. Nodes *i* and *j* have got a weighted connection if vectors x_i and x_j are close, according to an arbitrary measure of similarity. There are many ways to build an adjacency graph. The *cosine* graph with cosine weighting scheme is explored: given two vectors x_i and x_j, the weight w_{ij} between them is set by

$$w_{ij} = max\{0, \frac{cos(x_i, x_j) - \tau}{|cos(x_i, x_j) - \tau|} \cdot cos(x_i, x_j)\} \quad (1)$$

where a cosine threshold τ is necessary. The adjacency graph can be represented by using a symmetric $m \times m$ adjacency matrix, named W, whose element W_{ij} contains the weight between nodes *i* and *j*. The method of constructing an adjacency graph outlined above is correct if the data actually lie on a low dimensional manifold. Once such an adjacency graph is obtained, LPP will try to optimally preserve it in choosing projections.

Solve an Eigenmap problem. Compute the eigenvectors and eigenvalues for the generalized eigenvector problem:

$$XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}$$

where X is a $n \times m$ matrix whose columns are the original m vectors in R^n, D is a diagonal $m \times m$ matrix whose entries are column (or row) sums of W, $D_{ii} = \sum_j W_{ij}$ and $L = D - W$ is the Laplacian matrix. The solution of this problem is the set of eigenvectors $\mathbf{a}_0, \mathbf{a}_1, .., \mathbf{a}_{n-1}$, ordered according to their eigenvalues $\lambda_0 < \lambda_1 < .. < \lambda_{n-1}$. LPP projection matrix A is obtained by selecting the k eigenvectors corresponding to the k smallest eigenvalues: therefore it is a $n \times k$ matrix whose columns are the selected n-dimensional k eigenvectors. Final projection of original vectors into R^k can be linearly performed by $Y = A^T X$. This transformation provides a valid kernel that can be efficiently embedded into a classifier.

Embedding predicate knowledge through LPPs. While LSA finds a projection, according to the global properties of the space, LPP tries to preserve the local structures of the data. LPP exploits the adjacency graph in order to represent neighborhood information. It computes a transformation matrix which maps data points into a lower dimensional subspace. As the construction of an adjacency graph G can be based on any principle, its definition could account on some external information reflecting prior knowledge available about the task.

In this work, prior knowledge about LUs is embedded by exploiting their membership to frame dictionaries, thus removing from the graph all connections between LUs x_i and x_j that do not evoke the same prototypical situation. More formally Equation 1 can be rewritten more formally as:

$$w_{ij} = max\{0, \frac{cos(x_i, x_j) - \tau}{|cos(x_i, x_j) - \tau|} \cdot cos(x_i, x_j) \cdot \delta(i, j)\}$$

where

$$\delta(i,j) = \begin{cases} 1 & \textbf{iff } \exists F \; s.t. \; LU_i \in F \wedge LU_j \in F \\ 0 & \text{otherwise} \end{cases}$$

so the resulting manifold keeps close all LUs evoking the same frame. Since the number of connections could introduce too many constraints to the Eigenmap problem, a threshold is introduced to avoid the space collapse: for each LU, only the most-similar c connections are selected. The adoption of the proper *a priori* knowledge about the target task can be thus seen as a promising research direction.

4 Empirical Analysis

In this section the empirical evaluation of distributional models applied to the task of inducing LUs is presented. Different spaces obtained through the dimensionality reduction techniques imply different kernel functions used to independently train different SVMs. Our aim is to investigate the impact of these kernels in capturing both the frames and LUs' properties, as well as the effectiveness of their possible combination.

The problem of LUs' induction is here treated as a multi-classification problem, where each LU is considered as a positive or negative instance of a frame. We use Support Vector Machines (SVMs), (Joachims, 1999) a maximum-margin classifier that realizes a linear discriminative model. In case of not linearly separable examples, convolution functions $\phi(\cdot)$ can be used in order to transform the initial feature space into another one, where a hyperplane that separates the data with the widest margin can be found. Here new similarity measures, the kernel functions, can be defined through the dot-product $K(o_i, o_j) = \langle \phi(o_i) \cdot \phi(o_j) \rangle$ over the new representation. In this way, kernel functions K_{LSA} and K_{LPP} can be induced through the dimensionality reduction techniques ϕ_{LSA} and ϕ_{LPP} respectively, as described in sections 3.2 and 3.3. Kernel methods are advantageous because the combination of of kernel functions can be integrated into the SVM as they are still kernels. Consequently, the kernel combination $\alpha K_{LSA} + \beta K_{LPP}$ linearly combines the global properties captured by LSA and the locality constraints imposed by the LPP transformation. Here, parameters α and β weight the combination of the two kernels. The evoking frame for a novel LU is the one whose corresponding SVM has the highest (possibly negative) margin, according to a *one-*

	train	tune	test	overall
max	107	35	34	176
avg	28	8	8	44
total	2466	722	723	3911

Table 1: Number of LU examples for each data set from the 100 frames

vs-all scheme. In order to evaluate the quality of the presented models, accuracy is measured as the percentage of LUs that are correctly re-assigned to their original (gold-standard) frame. As the system can suggest more than one frame, different accuracy levels can be obtained. A LU is *correctly assigned* if its correct frame (according to FrameNet) belongs to the set of the best b proposals by the system (i.e. the first b scores from the underlying SVMs). Assigning different values to b, we obtained different levels of accuracy as the percentage of LUs that is correctly assigned among the first b proposals, as shown in Table 3.

4.1 Experimental Setup

The adopted gold standard is a subset of the FrameNet database and it consists of the most 100 represented frames in term of annotated examples and LUs. As the number of example is extremely unbalanced across frames[1], the LUs dictionary of each selected frame contains at least 10 LUs. It is a reasonable amount of information for the SVMs training and it is still a representative data set, being composed of 3,911 LUs, i.e. the 55% of the entire dictionary[2] of 7,230 evoking words. All word spaces are derived from the British National Corpus (BNC), which is underlying FrameNet and consisting of about 100 million words for English. Each selected frame is represented into the BNC by at least 362 annotated sentences, as the lack of a reasonable number of examples hardly produces a good distributional model of LUs. Each frame's list of LUs is split into train (60%), tuning (20%) and test set (20%) and LUs having Part-of-speech different from verb, noun or adjective are removed. In Table 1 the number of LUs for each set, as well as the maximum and the average number per frame, are summarized.

Four different approaches for the Word Space construction are used. The first two correspond to a Word-Based space, the last to a Syntax-Based, as described in section 3.1:

Window-*n* (**W*n***): contextual features correspond to the set of the 20,000 most frequent lemmatized words in the BNC. The association measure between LUs and contexts is the Point-wise Mutual Information (PMI). Valid contexts for LUs are fixed to a n-window. Hereafter two window width values will be investigated: *Window5* (*W5*) and *Window10* (*W10*).

Sentence (**Sent**): contextual features are the same above, but the valid contexts are extended to the entire sentence length.

SyntaxBased (**SyntB**): contextual features have been computed according to the "dependency-based" vector space discussed[3] in (Pado and Lapata, 2007). Observable contexts here are made of syntactically-typed co-occurrences within dependency graphs built from the entire set of BNC sentences. The most frequent 20,000 basic features, i.e. (syntactic relation,lemma) pairs, have been employed as contextual features corresponding to PMI scores. Syntactic relations are extracted using the Minipar parser.

Word space models thus focus on the LUs of the selected 100 frames and the dimensionality have been reduced by applying LSA and LPP at a new size of $l = 100$. Any prior knowledge information is provided to the tuning and test sets during the LPP transformation: the construction of the reduced feature space takes in account only LUs from the train set while remaining predicates are represented through the LPP linear projection. In these experiments the cosine threshold τ and the maximum number of constraints c are estimated over the tuning set and the best parametrizations are shown in Table 2. The adopted implementation of SVM is SVM-Light-TK [4].

4.2 Results

In these experiments the impact of the lexical knowledge gathered by different word-spaces is evaluated over the LU induction task. Moreover, the improvements achieved through LSA and LPP is measured. SVM classifiers are trained over the semantic spaces produced through the dimension-

[1]For example the SELF_MOTION frame counts 6,248 examples while 119 frames are represented by less than 10 examples

[2]The entire database contains 10,228 LUs and the number of evoking word is 7,230, without taking in account multiple frame assignments.

[3]The `Minimal` context provided by the Dependency Vectors tool is used. It is available at http://www.nlpado.de/∼sebastian/dv.html

[4]SVM-Light-TK is available at the url http://disi.unitn.it/∼moschitt/Tree-Kernel.htm

	α/β											τ	c
	1.0/0.0	.9/.1	.8/.2	.7/.3	.6/.4	.5/.5	.4/.6	.3/.7	.2/.8	.1/.9	0.0/1.0		
$W5$	0.668	0.669	0.672	**0.673**	0.669	0.662	0.649	0.632	0.612	0.570	0.033	0.55	5
$W10$	0.615	**0.619**	0.618	0.612	0.604	0.597	0.580	0.575	0.565	0.528	0.048	0.65	3
$Sent$	0.557	0.567	0.580	**0.584**	0.574	0.564	0.561	0.545	0.523	0.496	0.048	0.80	5
$SyntB$	0.654	0.**664**	0.662	0.652	0.651	0.647	0.649	0.634	0.627	0.592	0.056	0.40	3

Table 2: Accuracy at different combination weights of kernel $\alpha K_{LSA} + \beta K_{LPP}$ (specific baseline is 0.043)

	b-1	b-2	b-3	b-4	b-5	b-6	b-7	b-8	b-9	b-10	α/β
$W5_{orig}$	0,563	0,685	0,733	0,770	0,801	0,835	0,841	0,854	0,868	0,879	-
$W10_{orig}$	0,510	0,634	0,707	0,776	0,810	0,830	0,841	0,857	0,865	0,875	-
$Sent_{orig}$	0,479	0,618	0,680	0,734	0,764	0,793	0,813	0,837	0,845	0,852	-
$SyntB_{orig}$	0,585	0,741	0,803	0,840	0,866	0,874	0,886	0,903	0,907	0,913	-
$W5_{LSA+LPP}$	**0.673**	0.781	0.831	**0.865**	**0.881**	0.891	**0.906**	**0.912**	**0.926**	**0.938**	0.7/0.3
$W10_{LSA+LPP}$	0.619	0.739	0.786	0.818	0.849	0.865	0.878	0.888	0.901	0.909	0.9/0.1
$Sent_{LSA+LPP}$	0.584	0.705	0.766	0.798	0.825	0.835	0.848	0.864	0.876	0.889	0.7/0.3
$SyntB_{LSA+LPP}$	0.664	**0.791**	**0.840**	0.864	0.878	**0.893**	0.901	0.903	0.907	0.911	0.9/0.1

Table 3: Accuracy of original word-space models (*orig*) and semantic space models (*LSA+LPP*) on best-k proposed frames

ality reduction transformations. Representations of both semantic spaces are linearly combined as $\alpha K_{LSA} + \beta K_{LPP}$, where kernel weights α and β are estimated over the tuning set. Both kernels are used even without a combination: a ratio $\alpha = 1.0/\beta = 0.0$ denotes the LSA kernel alone, while $\alpha = 0.0/\beta = 1.0$ the LPP kernel. Table 2 shows best results, obtained through a RBF kernel. The $Window5$ model achieves the highest accuracy, i.e. 67% of correct classification, where a baseline of 4.3% is estimated assigning LUs to the most likely frame in the training set (i.e. the one containing the highest number of LUs). Wider windows achieve lower classification accuracy confirming that most of lexical information tied to a frame is near the LU. The Syntactic-based word space does not outperform the accuracy of a word-based space. The combination of both kernels has always provided the best outcome and the LSA space seems to be more accurate and expressive respect to the LPP one, as shown in Figure 1. In particular LPP alone is extremely unstable, suggesting that constraints imposed by the prior knowledge are orthogonal with respect to the corpus statistics.

Further experiments are carried out using the original co-occurrence space models, to assess improvements due to LSA and LPP kernel. In the latter investigation linear kernel achieved best results as confirmed in (Bengio et al., 2005), where the sensitivity to the curse of dimensionality of a large class of modern learning algorithms (e.g.

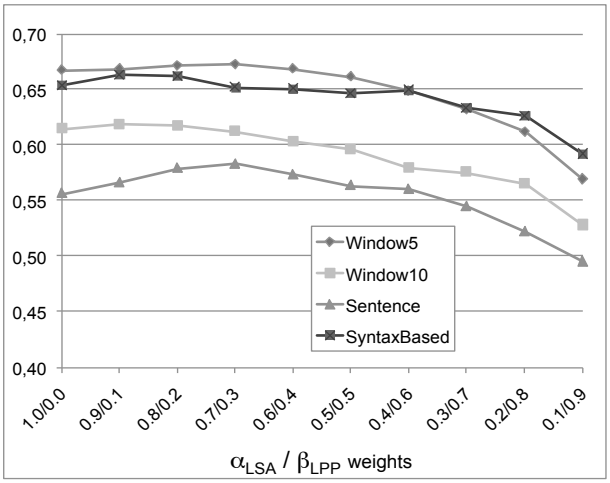

Figure 1: Accuracy at different combination weights of kernel $\alpha K_{LSA} + \beta K_{LPP}$

SVM) based on local kernels (e.g. RBF) is argued. As shown in Table 3, the performance drop of original (*orig*) models against the best kernel combination of LSA and LPP are significant, i.e. $\sim$ 10%, showing how the latent semantic spaces better capture properties of frames, avoiding data-sparseness, dimensionality problem and low-regularities of data-distribution.

Moreover, Table 3 shows how the accuracy level largely increases when more than one frame is considered: at a level $b = 3$, i.e. the novel LU is correctly classified if one of the original frames is comprised in the list (of three frames) proposed by the system, accuracy is 0.84 (i.e the SyntaxBased model), while at $b = 10$ accuracy is

LU (# WN_{syns})	frame_1	frame_2	frame_3	Correct frames
boil.v (5)	FOOD	FLUIDIC_MOTION	CONTAINERS	CAUSE_HARM
clap.v (7)	SOUNDS	MAKE_NOISE	COMMUNICATION_NOISE	BODY_MOVEMENT
crown.n (12)	LEADERSHIP	**ACCOUTREMENTS**	PLACING	ACCOUTREMENTS OBSERVABLE_BODYPARTS
school.n (7)	**EDUCATION_TEACHING**	BUILDINGS	**LOCALE_BY_USE**	EDUCATION_TEACHING LOCALE_BY_USE AGGREGATE
threat.n (4)	HOSTILE_ENCOUNTER	IMPACT	**COMMITMENT**	COMMITMENT
tragedy.n (2)	**TEXT**	KILLING	EMOTION_DIRECTED	TEXT

Table 4: Proposed 3 frames for each LU (ordered by SVM scores) and correct frames provided by the FrameNet dictionary. In parenthesis the number of different WordNet lexical senses for each LU.

nearly 0.94 (i.e Window5). It is high enough to support tasks such as the semi-automatic creation of new FrameNets. An error analysis indicates that many misclassifications are induced by a lack in the frame annotations, especially those concerning polysemic LUs[5]. Table 4 reports the analysis of a LU subset where the first 3 frames proposed for each evoking word are shown, ranked by the margin of the SMVs. The last column contains the frames evoked by LUs, according to the FrameNet dictionary, and the frame names in bold suggest their correct classification. Some LUs, like $threat$ (characterized by 4 lexical senses) seem to be misclassified: in this case the FrameNet annotation regards a specific sense that evokes the COMMITMENT frame (e.g. "There was a real *threat* that she might have to resign") without taking in account other senses like WordNet's "menace, threat (something that is a source of danger)" that could evoke the HOSTILE_ENCOUNTER frame. In other cases proposed frames seem to enrich the LUs dictionary, like BUILDINGS, here evoked by *school*.

5 Conclusions

The core purpose of this was to present an empirical investigation of the impact of different distributional models on the lexical unit induction task. The employed word-spaces, based on different co-occurrence models (either context and syntax-driven), are used as vector models of the LU semantics. On these spaces, two dimensionality reduction techniques have been applied. Latent Semantic Analysis (LSA) exploits global properties of data distributions and results in a global model for lexical semantics. On the other hand, the Locality Preserving Projection (LPP) method, that exploits regularities in the neighborhood of each lexical predicate, is also employed in a semi-supervised manner: local constraints expressing prior knowledge on frames are defined in the adjacency graph. The resulting embedding is therefore expected to determine a new space where regions for LU of a given frame can be more easily discovered. Experiments have been run using the resulting spaces for task dependent kernels in a SVM learning setting. The application of the FrameNet KB on the 100 best represented frames showed that a combined use of the global and local models made available by LSA and LPP, respectively, achieves the best results, as the 67.3% of LUs recovers the same frames of the annotated dictionary. This is a significant improvement with respect to previous results achieved by the pure distributional model reported in (Pennacchiotti et al., 2008).

Future work is required to increase the level of constraints made available from the semi-supervised setting of LPP: syntactic information, as well as role-related evidence, can be both accommodated by the adjacency constraints imposed for LPP. This constitutes a significant area of research towards a comprehensive semi-supervised model of frame semantics, entirely based on manifold learning methods, of which this study on LSA and LPP is just a starting point.

Acknowledgement We want to acknowledge Prof. Roberto Basili because this work would not exist without his ideas, inspiration and invaluable support.

[5] According to WordNet, in our dataset an average of 3.6 lexical senses for each LU is estimated.

References

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of SemEval-2007*,

pages 99–104, Prague, Czech Republic, June. Association for Computational Linguistics.

Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond Mooney. 2006. Probabilistic semi-supervised clustering with constraints. In *Semi-Supervised Learning*, pages 73–102. MIT Press.

Yoshua Bengio, Olivier Delalleau, and Nicolas Le Roux. 2005. The curse of dimensionality for local kernel machines. Technical report, Departement d'Informatique et Recherche Operationnelle.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet Detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*. Peter Lang, Frankfurt/Main.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *In Proceedings of ICML '08*, pages 160–167, New York, NY, USA. ACM.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining word sense and usage for modeling frame semantics. In *In Proceedings of STEP 2008, Venice, Italy.*

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 4(2):222–254.

G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proc. of SIGIR '88*, New York, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Yoav Goldberg and Michael Elhadad. 2009. On the role of lexical features in sequence labeling. In *In Proceedings of EMNLP '09*, pages 1142–1151, Singapore.

Zellig Harris. 1964. Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, *The Philosophy of Linguistics*, New York. Oxford University Press.

Xiaofei He and Partha Niyogi. 2003. Locality preserving projections. In *Proceedings of NIPS03*, Vancouver, Canada.

T. Joachims. 1999. *Making large-Scale SVM Learning Practical*. MIT Press, Cambridge, MA.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, Tartu, Estonia, May 24.

Richard Johansson and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of COLING*, Manchester, UK, August 18-22.

Tom Landauer and Sue Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104.

Dekang Lin. 1998. Automatic retrieval and clustering of similar word. In *Proceedings of COLING-ACL*, Montreal, Canada.

Alessandro Moschitti, Paul Morarescu, and Sanda M. Harabagiu. 2003. Open domain information extraction via automatic semantic labeling. In *FLAIRS Conference*, pages 397–401.

Sebastian Pado and Katrin Erk. 2005. To cause or not to cause: Cross-lingual semantic matching for paraphrase modelling. In *Proceedings of the Cross-Language Knowledge Induction Workshop*, Cluj-Napoca, Romania.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of framenet lexical units. In *Proceedings of The Empirical Methods in Natural Language Processing (EMNLP 2008) Waikiki, Honolulu, Hawaii*.

S.T. Roweis and L.K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University.

G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18:613Äì620.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL*, pages 12–21, Prague.

Mihai Surdeanu, , Mihai Surdeanu, A Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *In Proceedings of ACL 2003*.

Marta Tatu and Dan I. Moldovan. 2005. A semantic approach to recognizing textual entailment. In *HLT/EMNLP*.

J. B. Tenenbaum, V. Silva, and J. C. Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323.

Xin Yang, Haoying Fu, Hongyuan Zha, and Jesse Barlow. 2006. Semi-supervised nonlinear dimensionality reduction. In *23rd International Conference on Machine learning*, pages 1065–1072, New York, NY, USA. ACM Press.

Zhenyue Zhang and Hongyuan Zha. 2004. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Scientific Computing*, 26(1):313–338.