Learning Kernels for Semantic Clustering: A Deep Approach

Ignacio Arroyo-Fernández
Universidad Nacional Autónoma de México (UNAM)
iarroyof@iingen.unam.mx

Abstract

In this thesis proposal we present a novel semantic embedding method, which aims at consistently performing semantic clustering at sentence level. Taking into account special aspects of *Vector Space Models (VSMs)*, we propose to learn *reproducing kernels* in classification tasks. By this way, capturing spectral features from data is possible. These features make it theoretically plausible to model *semantic similarity criteria* in Hilbert spaces, i.e. the embedding spaces. We could improve the semantic assessment over embeddings, which are criterion-derived representations from traditional semantic vectors. The learned kernel could be easily *transferred* to clustering methods, where the Multi-Class Imbalance Problem is considered (e.g. semantic clustering of definitions of terms).

1 Introduction

Overall in Machine Learning algorithms (Duda et al., 2012), knowledge is statistically embedded via the Vector Space Model (VSM), which is also named *the semantic space* (Landauer et al., 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010). Contrarily to it is usually conceived in text data analysis (Manning et al., 2009; Aggarwal and Zhai, 2012), not any data set is suitable to embed into ℓ_p metric spaces, including euclidean spaces ($p = 2$) (Riesz and Nagy, 1955). This implies that, in particular, clustering algorithms are being adapted to some ℓ_p-derived metric, but not to semantic vector sets (clusters) (Qin et al., 2014).

The above implication also means that semantic similarity measures are commonly not consistent, e.g. the cosine similarity or transformation-based distances (Sidorov et al., 2014). These are mainly based on the concept of triangle. Thus if the triangle inequality does not hold (which induces norms for *Hilbert spaces* exclusively), then the case of the cosine similarity becomes mathematically inconsistent[1]. Despite VSMs are sometimes not mathematically analyzed, traditional algorithms work well enough for global semantic analysis (hereinafter *global analysis*, i.e. at document level where Zipf's law holds). Nevertheless, for local analysis (hereinafter *local analysis*, i.e., at sentence, phrase or word level) the issue remains still open (Mikolov et al., 2013).

In this thesis proposal, we will address the main difficulties raised from traditional VSMs for local analysis of text data. We consider the latter as an ill-posed problem (which implies unstable algorithms) in the sense of some explicit *semantic similarity criterion* (hereinafter *criterion*), e.g. topic, concept, etc. (Vapnik, 1998; Fernandez et al., 2007). The following feasible reformulation is proposed. By learning a kernel in classification tasks, we want to induce an embedding space (Lanckriet et al., 2004; Cortes et al., 2009). In this space, we will consider relevance (weighting) of spectral features of data, which are in turn related to the shape of semantic vector sets (Xiong et al., 2014). These vectors would be derived from different *Statistical Language Models (SLMs)*; i.e. countable things, e.g. *n*-grams, bag-of-words (BoW), etc.; which in turn encode *language*

[1] Riesz (1955) gives details about Hilbert spaces.

Proceedings of NAACL-HLT 2015 Student Research Workshop (SRW), pages 79–87,
Denver, Colorado, June 1, 2015. ©2015 Association for Computational Linguistics

aspects (e.g. semantics, syntax, morphology, etc.). Learned kernels are susceptible to be transferred to clustering methods (Yosinski et al., 2014; Bengio et al., 2014), where spectral features would be properly *filtered* from text (Gu et al., 2011).

When both learning and clustering processes are performed, the kernel approach is tolerant enough for data scarcity. Thus, eventually, we could have any criterion-derived amount of semantic clusters regardless of the Multi-Class Imbalance Problem (MCIP) (Sugiyama and Kawanabe, 2012). It is a rarely studied problem in Natural Language Processing (NLP), however, contributions can be helpful in a number of tasks such as IE, topic modeling, QA systems, opinion mining, Natural Language Understanding, etc.

This paper is organized as follows: In Section 2 we show our case study. In Section 3 we show the embedding framework. In Section 4 we present our learning problem. Sections 5 and 6 respectively show research directions and related work. In Section 7, conclusions and future work are presented.

2 A case study and background

A case study. Semantic clustering of definitions of terms is our case study. See the next extracted[2] examples for the terms *window* and *mouse*. For each of them, the main acception is showed first, and afterwards three secondary acceptions:

1. *A **window** is a **frame** including a **sheet of glass** or other material capable of admitting light...*
   (a) *The window is the **time** elapsed since a **passenger** calls to **schedule**...*
   (b) *A window is a **sequence** region of 20-codon length on an alignment of **homologous genes**...*
   (c) *A window is any **GUI element** and is usually identified by a Windows handle...*
2. *A **mouse** is a **mammal** classified in the order **Rodentia**, suborder **Sciurognathi**....*
   (a) *A mouse **is a small object** you can roll along a hard, flat surface...*
   (b) *A mouse is a **handheld pointing device** used to position a cursor on a **computer**...*
   (c) *The Mouse is a **fictional character** in **Alice's Adventures** in Wonderland by **Lewis Carroll**...*

In the example 1, it is possible to assign the four acceptions to four different semantic groups (the window (1), transport services (1a), genetics (1b) and computing (1c)) by using lexical features (bold terms). This example also indicates how abstract concepts are always latent in the definitions. The example 2 is a bit more complex. Unlike to example 1, there would be three clusters because there are two semantically similar acceptions (2a and 2b are related to computing). However, they are lexically very distant. See that in both examples the amount of semantic clusters can't be defined a priory (unlike to Wikipedia). Additionally, it is impossible to know what topic the users of an IE system could be interested in. These issues, point out the need for analyzing the way we are currently treating semantic spaces in the sense of stability of algorithms (Vapnik, 1998), i.e. the existence of semantic similarity consistence, although Zipf's law scarcely holds (e.g. in local analysis).

Semantic spaces and embeddings. Erk (2012) and Brychcín (2014) showed insightful empiricism about well known semantic spaces for different cases in global analysis. In this work we have special interest in local analysis, where semantic vectors are representations (*embeddings*) derived from learned feature maps for specific semantic assessments (Mitchell and Lapata, 2010). These feature maps are commonly encoded in Artificial Neural Networks (ANNs) (Kalchbrenner et al., 2014).

ANNs have recently attracted worldwide attention. Given their surprising adaptability to unknown distributions, they are used in NLP for embedding and feature learning in local analysis, i.e. *Deep Learning* (DL) (Socher et al., 2011; Socher et al., 2013). However, we require knowledge transfer towards clustering tasks. It is still not feasible by using ANNs (Yosinski et al., 2014). Thus, theoretical access becomes ever more necessary, so it is worth extending *Kernel Learning* (KL) studies as alternative feature learning method in NLP (Lanckriet et al., 2004). Measuring subtle semantic displacements, according to a criterion, is theoretically attainable in a well defined (learned) *reproducing kernel Hilbert space* (RKHS), e.g. some subset of L_2 (Aronszajn, 1950). In these spaces, features are latent *abstraction levels*[3] of data spectrum, which improves kernel scaling (Dai et al., 2014; Anandkumar et al., 2014).

[2] `www.describe.com.mx`

[3] Mainly in *DL*, it is known there are different hierarchies of generality of features learned by a learning machine.

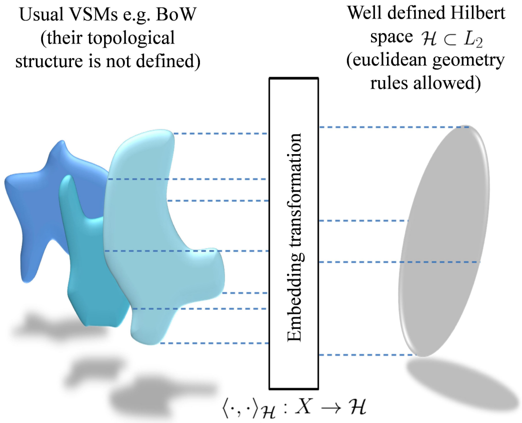
Usual VSMs e.g. BoW (their topological structure is not defined)
Well defined Hilbert space $\mathcal{H} \subset L_2$ (euclidean geometry rules allowed)
Embedding transformation
$\langle\cdot,\cdot\rangle_{\mathcal{H}} : X \to \mathcal{H}$

Figure 1: General schema of the transformation framework from some traditional VSM (left) to a well defined embedding space (right).

3 RKHS and semantic embeddings

We propose mapping sets of semantic vectors (e.g. BoW) into well defined function spaces (RKHSs), prior to directly endowing such sets (not elliptical or at least convex (Qin et al., 2014)) with the euclidean norm, $\|.\|_2$ (see Figure 1). For the aforesaid purpose, we want to take advantage of the RKHSs.

Any semantic vector $x_o \in X$ could be consistently embedded (*transformed*) into a well defined Hilbert space by using the *reproducing property* of a kernel $k(\cdot,\cdot)$ (Shawe-Taylor and Cristianini, 2004):

$$f_{x_o}(x) = \langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}}; \;\; \forall x \in X \qquad (1)$$

where: $\mathcal{H} \subset L_2$ is a RKHS, $f_{x_o}(\cdot) \in \mathcal{H}$ is the embedding derived from x_o, which can be seen as fixed parameter of $k(\cdot, x_o) = f(\cdot) \in \mathcal{H}$. This embedding function is defined over the vector domain $\{x\} \subset X$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}} : X \to \mathcal{H}$ is the inner product in $\mathcal{H}$.

Always that (1) holds, $k(\cdot,\cdot)$ is a positive definite (PD) kernel function, so X does not need even to be a vector space and even then, convergence of any sequence $\{f_n(x) : f_n \in \mathcal{H}; n \in \mathbb{N}\}$ can be ensured. The above is a highly valuable characteristic of the resulting function space (Smola et al., 2007):

$$\lim_{n\to\infty} f_n = f \iff \lim_{n\to\infty} k_n(\cdot, x) = k(\cdot, x). \quad (2)$$

The result (2) implies that convergence of summation of initial guessing kernel functions $k_n(\cdot,\cdot) \in \mathcal{H}$ always occurs, hence talking about the existence of a suitable kernel function $k(\cdot,\cdot) \in \mathcal{H}$ in (1) is absolutely possible. It means that L_2 operations can be consistently applied, e.g. the usual norm $\|\cdot\|_2$, trigonometric functions (e.g. $\cos\theta$) and distance $d_2 = \|f_n - f_m\|_2 : m \neq n$. Thus, from right side of (2), in order that (1) holds convergence of the Fourier series decomposition of $k(\cdot,\cdot)$ towards the spectrum of desired features from data is necessary; i.e., by learning parameters and hyperparameters[4] of the series (Ong et al., 2005; Băzăvan et al., 2012).

3.1 Learnable kernels for language features

Assume (1) and (2) hold. For some SLM a encoded in a traditional semantic space, it is possible to define a learnable kernel matrix K_a as follows (Lanckriet et al., 2004; Cortes et al., 2009):

$$K_a := \sum_{i=1}^{p} \beta_i K_i, \qquad (3)$$

where $\{K_i\}_{i=1}^p \subset \mathcal{K}$ is the set of p initial guessing kernel matrices (belonging to the family $\mathcal{K}$, e.g. Gaussian) with fixed hyperparameters and β_i's are parameters weighting K_i's. Please note that, for simplicity, we are using matrices associated to kernel functions $k_i(\cdot,\cdot), k_a(\cdot,\cdot) \in \mathcal{H}$, respectively.

In the Fourier domain and bandwidth. In fact (3) is a Fourier series, where β_i's are decomposition coefficients of K_a (Băzăvan et al., 2012). This kernel would be fitting the spectrum of some SLM that encodes some latent language aspect from text (Landauer et al., 1998). On one hand, in Fourier domain operations (e.g. the error vector norm) are closed in L_2, i.e., according to (2) convergence is ensured as a Hilbert space is well defined. Moreover, the L_2-regularizer is convex in terms of the Fourier series coefficients (Cortes et al., 2009). The aforementioned facts imply benefits in terms of computational complexity (scaling) and precision (Dai et al., 2014). On the other hand, hyperparameters of initial guessing kernels are learnable for detecting the bandwitdh of data (Ong et al., 2005; Băzăvan et al., 2012; Xiong et al., 2014). Eventually, the latter fact would lead us to know (learning) bounds for

[4] So called in order to make distinction between weights (kernel parameters or coefficients) and the basis function parameters (hyperparameters), e.g. mean and variance.

the necessary amount of data to properly train our model (*the Nyquist theorem*).

Cluster shape. A common shape among clusters is considered even for unseen clusters with different, independent and imbalanced prior probability densities (Vapnik, 1998; Sugiyama and Kawanabe, 2012). For example, if data is Guassian-distributed in the input space, then shape of different clusters tend to be elliptical (the utopian ℓ_2 case), although their densities are not regular or even very imbalanced. Higher abstraction levels of the data spectrum possess mentioned traits (Ranzato et al., 2007; Baktashmotlagh et al., 2013). We will suggest below a more general version of (3), thereby considering higher abstraction levels of text data.

4 Learning our kernel in a RKHS

A transducer is a setting for learning parameters and hyperparameters of a multikernel linear combination like the Fourier series (3) (Băzăvan et al., 2012).

Overall, the above setting consists on defining a multi-class learning problem over a RKHS: let $\mathcal{Y}_\theta = \{y_\ell\}_{y_\ell \in \mathbb{N}}$ be a sequence of targets inducing a semantic criterion θ, likewise a training set $\mathcal{X} = \{x_\ell\}_{x_\ell \in \mathbb{R}^n}$ and a set of initial guessing kernels $\{K_{\sigma_i}\}_{i=1}^p \subset \mathcal{K}$ with the associated hyperparameter vector $\sigma_a = \{\sigma_i\}_{i=1}^p$. Then for some SLM $a \in \mathcal{A}$, we would learn the associated kernel matrix K_a by optimizing the SLM empirical risk functional:

$$\mathcal{J}_{\mathcal{A}}(\sigma_a, \beta_a) = L_{\mathcal{A}}(K_a, \mathcal{X}, \mathcal{Y}_\theta) + \psi(\sigma_a) + \xi(\beta_a), \tag{4}$$

where in $\mathcal{J}_{\mathcal{A}}(\cdot, \cdot)$ we have:

$$K_a = \sum_{1 \le i \le p} \beta_i K_{\sigma_i}. \tag{5}$$

The learning is divided in two interrelated stages: at the first stage, the free parameter vector $\beta_a = \{\beta_i\}_{i=1}^p$ in (5) (a particular version of (3)), is optimized for learning a partial kernel $\widehat{K}_a$, given a fixed (sufficiently small) σ_a and by using the regularizer $\xi(\beta_a)$ over the SLM prediction loss $L_{\mathcal{A}}(\cdot, \cdot)$ in (4). Conversely at the second stage σ_a is free, thus by using the regularizer $\psi(\sigma_a)$ over the prediction loss $L_{\mathcal{A}}(\cdot, \cdot)$, given that the optimal β_a^* was found at the first stage, we could have the optimal σ_a^* and therefore K_a^* is selected.

At higher abstraction levels, given the association $\{\mathcal{X}, \mathcal{Y}_\theta\}$, the transducer setting would learn a kernel function that fits a multi-class partition of X via summation of K_a's. Thus, we can use learned kernels K_a^* as new initial guesses in order to learn a compound kernel matrix K_θ for a higher abstraction level:

$$\mathcal{J}(\gamma_\theta) = L(K_\theta, \mathcal{X}, \mathcal{Y}_\theta) + \zeta(\gamma_\theta), \tag{6}$$

where in the general risk functional $\mathcal{J}(\cdot)$ we have:

$$K_\theta = \sum_{a \in \mathcal{A}} \gamma_a K_a^*. \tag{7}$$

In (6) the vector $\gamma_\theta = \{\gamma_a\}_{a \in \mathcal{A}}$ weights semantic representations K_a^* associated to each SLM and $\zeta(\gamma_\theta)$ is a proper regularizer over the general loss $L(\cdot, \cdot)$. The described learning processes can even be jointly performed (Băzăvan et al., 2012). The aforementioned losses and regularizers can be conveniently defined (Cortes et al., 2009).

4.1 The learned kernel function

In order to make relevant features to emerge from text, we would use our learned kernel K_θ^*. Thus if $\{\gamma_\theta^*, \{\beta_a^*, \sigma_a^*\}_{a \in \mathcal{A}}\}$ is the solution set of the learning problems (4) and (6), then combining (5) and (7) gives the embedding kernel function, for $|\mathcal{A}|$ different SLMs as required (see Figure 2):

Definition 1. *Given a semantic criterion θ, then the learned parameters $\{\gamma_\theta^*, \{\beta_a^*, \sigma_a^*\}_{a \in \mathcal{A}}\}$ are eigenvalues of kernels $\{K_a^*\}_{a \in \mathcal{A}} \prec K_\theta^*$, respectively*[5]. *Thus according to (1), we have for any semantic vector $x_o \in X$ its representation $f_{x_o}(x) \in \mathcal{H}$:*

$$\begin{aligned} f_{x_o}(x) &:= \sum_{a \in \mathcal{A}} \sum_{i=1}^{p} \gamma_a^* \beta_i^* k_i(x, x_o) \\ &= k_\theta(x, x_o) \approx K_\theta^* x_o. \end{aligned} \tag{8}$$

In (8), $k_i(\cdot, \cdot), k_\theta(\cdot, \cdot) \in \mathcal{H} \subset L_2$ are reproducing kernel functions associated to matrices K_{σ_i} and K_θ, respectively. The associated $\{\sigma_a^*\}_{a \in \mathcal{A}}$ would be optimally fitting the bandwidth of data. $X \supset \mathcal{X}$ is a compounding semantic space from different SLMs

[5] *(i)* The symbol '$\prec$' denotes subordination (from right to left) between operators, i.e. hierarchy of abstraction levels. *(ii)* See (Shawe-Taylor and Cristianini, 2004; Anandkumar et al., 2014) for details about eigendecompositions.

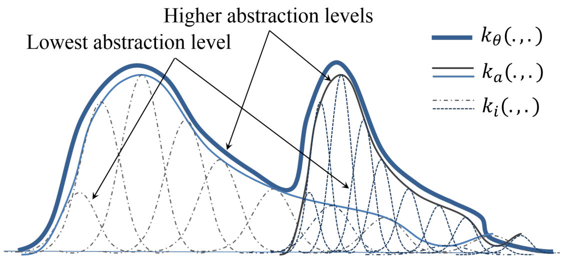

Figure 2: Sketch (bold plot) of the abstraction levels of some learned kernel function $k_\theta(\cdot,\cdot) \in \mathcal{H} \subset L_2$.

$a \in \mathcal{A}$ (Băzăvan et al., 2012). According to θ, semantic clustering could be consistently performed in $\mathcal{H}$ by computing any L_2 similarity measure between embeddings $\{f_{x_n}, f_{x_m}\}$, which are derived from any semantic vectors $x_n, x_m \in X$, e.g. *(i)* the kernel correlation coefficient $\rho_\theta = \mu k_\theta(x_n, x_m) \in [0, 1]$; with $\mu = \frac{1}{\|f_{x_n}\|\|f_{x_m}\|}$, and *(ii)* the distance by simply computing $d_2 = \|f_{x_n} - f_{x_m}\|_2$.

Please note that we could extend Definition 1 to deeper levels (layers) associated to abstraction levels of SLMs. These levels could explicitly encode morphology, syntax, semantics or compositional semantics, i.e. $\{K_a\}_{a \in \mathcal{A}} = K_{SLMs} \prec K_{aspects}$.

5 Research directions

Our main research direction is to address in detail linguistic interpretations associated to second member of (8), which is still not clear. There are potential ways of interpreting *pooling operations* over the expansion of either eigenvalues or eigenfunctions of $f_{x_o}(\cdot)$. This fact could lead us to an alternative way of analyzing written language, i.e. in terms of the spectral decomposition of $\mathcal{X}$ given θ.

As another direction we consider data scarcity (low annotated resources). It is a well handled issue by spectral approaches like the proposed one, so it is worth investigating hyperparameter learning techniques. We consider hyperparameters as the lowest abstraction level of the learned kernel and they are aimed at data bandwidth estimation (i.e. by tuning the σ_i associated to each $k_i(\cdot, \cdot)$ in (8)). This estimation could help us to try to answer the question of how much training data is enough. This question is also related to the quality bounds of a learned kernel. These bounds could be used to investigate the possible relation among the number of annotated clusters, the training set size and the generalization ability. The latter would be provided (transferred) by the learned kernel to a common clustering algorithm for discovering imbalanced unseen semantic clusters. We are planning to perform the above portrayed experiments at least for a couple of semantic criteria[6], including term acception discovering (Section 2). Nevertheless, much remains to be done.

6 Related work

Clustering of definitional contexts. Molina (2009) processed snippets containing definitions of terms (Sierra, 2009). The obtained PD matrix is not more than a homogeneous quadratic kernel that induces a Hilbert space: The *Textual Energy* of data (Fernandez et al., 2007; Torres-Moreno et al., 2010). Hierarchical clustering is performed over the resulting space, but some semantic criterion was not considered. Thus, such as Cigarran (2008), they ranked retrieved documents by simply relying on lexical features (global analysis). ML analysis was not performed, so their approach suffers from high sensibility to lexical changes (instability) in local analysis.

Paraphrase extraction from definitional sentences. Hashimoto, et.al. (2011) and Yan, et.al. (2013) engineered vectors from contextual, syntactical and lexical features of definitional sentence paraphrases (similarly to Lapata (2007) and Ferrone (2014)). As training data they used a POS annotated corpus of sentences that contain noun phrases. It was trained a binary SVM aimed at both paraphrase detection and multi-word term equivalence assertion (Choi and Myaeng, 2012; Abend et al., 2014). More complex constructions were not considered, but their feature mixure performs very well.

Socher et al., (2011) used ANNs for paraphrase detection. According to labeling, the network unsupervisedly capture as many language features as latent in data (Kalchbrenner et al., 2014). The network supervisedly learns to represent desired contents inside phrases (Mikolov et al., 2013); thus paraphrase detection is highly generalized. Nevertheless, it is notable the necessity of a tree parser. Unlike to (Socher et al., 2013), the network must to learn syntactic features separately.

[6] For example: SemEval-2014; Semantic Evaluation Exercises.

Definitional answer ranking. Fegueroa (2012) and (2014) proposed to represent definitional answers by a Context Language Model (CLM), i.e. a Markovian process as probabilistic language model. A knowledge base (WordNET) is used as an annotated corpus of specific domains (limited to Wikipedia). Unlike to our approach, queries must be previously disambiguated; for instance: *"what is a computer virus?"*, where "computer virus" disambiguates "virus". Answers are classified according to relevant terms (Mikolov et al., 2013), similarly to the way topic modeling approaches work (Fernandez et al., 2007; Lau et al., 2014).

Learning kernels for clustering. Overall for knowledge transfer from classification (source) tasks to clustering (target) tasks, the state of the art is not bast. This setting is generally explored by using toy Gaussian-distributed data and predefined kernels (Jenssen et al., 2006; Jain et al., 2010). Particularly for text data, Gu et.al. (2011) addressed the setting by using multi-task kernels for global analysis. In their work, it was not necessary neither to discover clusters nor to model some semantic criterion. Both them are assumed as a presetting of their analysis, which differs from our proposal.

Feasibility of KL over DL. We want to perform clustering over an embedding space. At the best of our knowledge there exist two dominant approaches for feature learning: KL and DL. However, knowledge transfer is equally important for us, so both procedures should be more intuitive by adopting the KL approach instead of DL. We show the main reasons: *(i) Interpretability*. The form (8) has been deducted from punctual items (e.g. SLMs encoding language aspects), which leads us to think that a latent statistical interpretation of language is worthy of further investigation. *(ii) Modularity*. Any kernel can be transparently transferred into kernelized and non-kernelized clustering methods (Schölkopf et al., 1997; Aguilar-Martin and De Mántaras, 1982; Ben-Hur et al., 2002). *(iii) Mathematical support*. Theoretical access provided by kernel methods would allow for future work on semantic assessments via increasingly abstract representations. *(iv) Data scarcity*. It is one of our principal challenges, so kernel methods are feasible because of their generalization predictability (Cortes and Vapnik, 1995).

Regardless of its advantages, our theoretical framework exhibit latent drawbacks. The main of them is that feature learning is not fully unsupervised, which suggests the underlying possibility of preventing learning from some decisive knowledge related to, mainly, the tractability of the MCIP. Thus, many empirical studies are pending.

7 Conclusions and future work

At the moment, our theoretical framework analyzes semantic embedding in the sense of a criterion for semantic clustering. However, correspondences between linguistic intuitions and the showed theoretical framework (interpretability) are actually incipient, although we consider these challenging correspondences are described in a generalized way in the seminal work of Harris (1968). It is encouraging (not determinant) that our approach can be associated to his operator hypothesis on composition and separability of both linguistic entities and language aspects. That is why we consider it is worth investigating spectral decomposition methods for NLP as possible rapprochement to elucidate improvements in semantic assessments (e.g. semantic clustering). Thus, by performing this research we also expect to advance the state of the art in statistical features of written language.

As immediate future work we are planning to learn compositional distributional *operators* (kernels), which can be seen as stable solutions of operator equations (Harris, 1968; Vapnik, 1998). We would like to investigate this approach for morphology, syntax and semantics (Mitchell and Lapata, 2010; Lazaridou et al., 2013). Another future proposal could be derived from the abovementioned approach (operator learning), i.e. multi-sentence compression for automatic sumarization.

A further extension could be ontology learning. It would be proposed as a multi-structure KL framework (Ferrone and Zanzotto, 2014). In this case, IE and knowledge organization would be our main aims (Anandkumar et al., 2014).

Aknowledgements. This work is funded by CONACyT Mexico (grant: 350326/178248). Thanks to the UNAM graduate program in CS. Thanks to Carlos Méndez-Cruz, to Yang Liu and to anonymous reviewers for their valuable comments.

References

Omri Abend, B. Shay Cohen, and Mark Steedman. 2014. Lexical inference over multi-word predicates: A distributional approach. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 644–654. ACL.

Charu C. Aggarwal and Cheng Xiang Zhai. 2012. An introduction to text mining. In Charu C Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 1–10. Springer US.

J Aguilar-Martin and R De Mántaras. 1982. The process of classification and learning the meaning of linguistic descriptors of concepts. *Approximate Reasoning in Decision Analysis*, 1982:165–175.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

Nachman Aronszajn. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. 2013. Unsupervised domain adaptation by domain invariant projection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 769–776. IEEE.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Eduard Gabriel Băzăvan, Fuxin Li, and Cristian Sminchisescu. 2012. Fourier kernel learning. In *Computer Vision–ECCV 2012*, pages 459–473. Springer.

Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. 2002. Support vector clustering. *The Journal of Machine Learning Research*, 2:125–137.

Yoshua Bengio, Ian J. Goodfellow, and Aaron Courville. 2014. Deep learning. Book in preparation for MIT Press.

Tomáš Brychcín and Miroslav Konopík. 2014. Semantic spaces for improving language modelling. *Computer Speech and Language*, 28:192–209.

Sung-Pil Choi and Sung-Hyon Myaeng. 2012. Terminological paraphrase extraction from scientific literature based on predicate argument tuples. *Journal of Information Science*, pages 1–19.

Juan Manuel Cigarrán Recuero. 2008. *Organización de resultados de búsqueda mediante análisis formal de conceptos*. Ph.D. thesis, Universidad Nacional de Educación a Distancia; Escuela Técnica Superior de Ingeniería Informática.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2009. L 2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press.

Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina F Balcan, and Le Song. 2014. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049.

Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.

Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Silvia Fernandez, Eric San Juan, and Juan-Manuel Torres-Moreno. 2007. Textual energy of associative memories: Performant applications of enertex algorithm in text summarization and topic segmentation. *MICAI 2007: Advances in Artificial Intelligence*, pages 861–871.

Lorenzo Ferrone and Fabio Massimo Zanzotto. 2014. Towards syntax-aware compositional distributional semantic models. In *Proceedings of COLING 2014: Technical Papers*, pages 721–730. Dublin City University and Association for Computational Linguistics (ACL).

Alejandro Figueroa and John Atkinson. 2012. Contextual language models for ranking answers to natural language definition questions. *Computational Intelligence*, pages 528–548.

Alejandro Figueroa and Günter Neumann. 2014. Category-specific models for ranking effective paraphrases in community question answering. *Expert Systems with Applications*, 41(10):4730–4742.

Quanquan Gu, Zhenhui Li, and Jiawei Han. 2011. Learning a kernel for multi-task clustering. In *Proceedings of the 25th AAAI conference on artificial intelligence*. Association for the Advancement of Artificial Intelligence (AAAI).

Zellig S. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1087–1097.

Prateek Jain, Brian Kulis, and Inderjit S Dhillon. 2010. Inductive regularized learning of kernel functions. In

Advances in Neural Information Processing Systems, pages 946–954.

Robert Jenssen, Torbjørn Eltoft, Mark Girolami, and Deniz Erdogmus. 2006. Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm. In *Advances in Neural Information Processing Systems*, pages 633–640.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. 2004. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, December.

Thomas K Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, volume 1, pages 259–270.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge UP.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(34):1388–1429. Cognitive Science Society, ISSN: 1551-6709.

A Molina. 2009. Agrupamiento semántico de contextos definitorios. *Mémoire de Master, Universidad Nacional Autónoma de México–Posgrado en Ciencia e Ingeniería de la Computación, México*, 108.

Cheng S Ong, Robert C Williamson, and Alex J Smola. 2005. Learning the kernel with hyperkernels. In *Journal of Machine Learning Research*, pages 1043–1071.

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

Danfeng Qin, Xuanli Chen, Matthieu Guillaumin, and Luc V Gool. 2014. Quantized kernel learning for feature matching. In *Advances in Neural Information Processing Systems*, pages 172–180.

M Ranzato, Fu Jie Huang, Y-L Boureau, and Yann LeCun. 2007. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

F. Riesz and Sz Nagy. 1955. Functional analysis. *Dover Publications, Inc., New York. First published in*, 3(6):35.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel principal component analysis. In *Artificial Neural Networks–ICANN'97*, pages 583–588. Springer.

Jhon Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge UP. ISBN: 978-0-521-81397-6.

Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3).

Gerardo Sierra. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *LinguaMÁTICA*, 2:13–38, Dezembro.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. 2007. A hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, pages 13–31. Springer-Verlag.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. Citeseer.

M. Sugiyama and M. Kawanabe. 2012. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. Adaptive computation and machine learning. MIT Press.

Juan-Manuel Torres-Moreno, Alejandro Molina, and Gerardo Sierra. 2010. La energía textual como medida de distancia en agrupamiento de definiciones. In *Statistical Analysis of Textual Data*, pages 215–226.

Vladimir Naumovich Vapnik. 1998. *Statistical learning theory*. Wiley New York.

Yuanjun Xiong, Wei Liu, Deli Zhao, and Xiaoou Tang. 2014. Zeta hull pursuits: Learning nonconvex data hulls. In *Advances in Neural Information Processing Systems*, pages 46–54.

Yulan Yan, Chikara Hashimoto, Kentaro Torisawa, Takao Kawai, Jun'ichi Kazama, and Stijn De Saeger. 2013. Minimally supervised method for multilingual paraphrase extraction from definition sentences on the web. In *HLT-NAACL*, pages 63–73.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328.