Distributional Lexical Semantics:  
towards uniform representation paradigms  
for advanced acquisition and processing tasks  

R. BASILI,  
University of Roma, Tor Vergata,  
Via della Ricerca Scientifica, 00133 Roma (Italy)  

M. PENNACCHIOTTI  
Yahoo! Inc., Santa Clara, USA  

(Received 31 May 2010; revised ....)  

Prologue  

“These long chains of perfectly simple and easy reasoning by means of which geometers  
are accustomed to carry out their most difficult demonstrations had led me to fancy that  
everything that can fall under human knowledge forms a similar sequence”  
(René Descartes, Discours de la Méthode, 1637)  

The distributional hypothesis states that words with similar distributional properties have similar semantic properties (Harris, 1968). This perspective on word semantics, early discussed in linguistics (Firth, 1957; Harris, 1968), constitutes today a building block of many lexical semantic researches in computational linguistics. Distributional methods are widely and successfully used for modelling different phenomena, ranging from sense disambiguation to textual inference.  

Distributional approaches, early explored in Information Retrieval (Salton et al., 1975), pushed for strongly lexicalized representations of text meaning. These simple meaning surrogates tackled complex problems with a surprising success, in terms of accuracy and scalability. Distributional notions (e.g. document frequency and word co-occurrence counts) have proved a key factor of success, as opposed to early logic-based approaches to relevance modeling in Information Retrieval (van Rijsbergen, 1986; Chiaramella and Chevallet, 1992; Van Rijsbergen and Lalmas, 1996).  

Although targets and tradition here differ substantially from Information Retrieval, these reasons of success must be taken into account for the development of a realistic research perspective in Natural Language Engineering. This special issue has the goal of presenting existing achievements, emphasizing potentials, and fostering forthcoming applications, in the field of distributional approaches for Natural Language Engineering.
1 Distributional Models in Language Learning and Processing

Large part of the “empirical renaissance” that since the early ’90s characterized most of the research in computational linguistics, is undoubtedly tied to a distributional view of language (Church and Mercer, 1993). In that period, the linguistic meaning of words was perceived as a side effect of the words’ distribution in large-scale texts collections. Typically, a word was described by a high-dimensional vector, which captures co-occurrence statistics of the words in the corpus. This results in mapping the original linguistic problem into a geometrical space, where distances between words’ vector are used as quantitative models of the linguistic problem. These spaces have been soon called semantic spaces, as the distributional hypothesis suggests (Lund and Burgess, 1997; Schutze, 1998).

Functions in these geometric spaces are today often advocated as representational, learning and processing models. As representational models, they allow to define and manipulate complex linguistic information, as, for example, real-valued vectors for individual lexical and text units. As learning models, they define effective ways of gathering, weighting, and mapping specific linguistic evidence. As processing models, geometric spaces allow to formalize accurate decision functions for supporting statistical inference\(^1\). Inductive methods proposed in corpus linguistics are indeed an embodiment of a distributional view on natural language and its semantics.

1.1 Distributional models for Linguistic Information Processing

Semantic spaces are usually built to capture contextual co-occurrences of words. They can be thus classified according to the types of such contexts. Different classes of contexts tend to emphasize different types of semantic properties, and thus semantic relations among words (Sahlgren, 2006).

In document-based spaces, contexts are entire documents. These spaces have been historically used in Information Retrieval, and tend of capture topic similarity between words – i.e. words similar in the space tend to refer to the same topic, as they appear in similar documents (e.g. ‘doctor’/‘hospital’) for the medical topic).

Word-based spaces (e.g., (Bullinaria and Levy, 2007)) define contexts as the words appearing in a \(n\)-window of the target (i.e. \(n\) tokens on the left and on the right). Such spaces model a generic notion of semantic relatedness. Two target words 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 association and phrasal association). Indeed, words with similar co-occurrences can be both words occurring together in a text (e.g. ‘doctor’ and ‘patient’ in “the doctor operated the patient”, sharing the same contexts ‘operated’) and substitutional words (e.g. ‘doctor’ and ‘surgeon’ in “the

\(^1\) Notice how Statistical Learning Theory (Vapnik, 1995) is just one of the recent result of a strictly geometrical view on the modeling of the inductive inference process.
(doctor-surgeon) operated the patient”). (Sahlgren, 2006) and (Erk and Padó, 2008) present interesting analysis and insights on these classes of semantic spaces.

More recently, (Pado and Lapata, 2007) introduced syntax-based spaces, where contexts correspond to lexical syntactic relations (e.g. X-VSubj-man for every target word X). These spaces are very effective in modeling semantic similarity. Two target words close in the space are likely to be in a paradigmatic relation, i.e. to be categorically close (as in a is-a hierarchy, (Budanitsky and Hirst, 2006; Lin, 1998)). Syntax-based spaces represent a natural and needed evolution of semantic space models. Indeed, it has been observed by many, that document-based and word-based spaces are often too simple to capture the structural complexity of texts. For the overall goal of understanding how meaning resides in texts, more complex linguistic features must to be taken into account: the study of syntax-based spaces represents a promising instantiation of this research line.

An open research challenge regards the attempt to better define how different distributional approaches can capture paradigmatic and syntagmatic properties. We know that different types of contexts emphasize either paradigmatic or syntagmatic properties, but their acquisition and use is also related to other factors: for example, the similarity measure available by a given semantic space. Symmetric similarity measures (e.g. distance metrics) are very important for acquiring and using paradigmatic classes, such as synonymy sets and taxonomic classes. On the contrary, antisymmetric relations and non commutative functions (e.g. geometrical projections into subspaces) seem more important for directional inferences (e.g. lexical generalizations or textual entailments).

Another promising area of research is the integration of different distributional models into a unifying one. Indeed, vector-based representations allow the unification of lexical representations emerging from different spaces, through the use of algebraic composition. This represents a relevant advantage with respect to other representation formalisms, but requires more in depth exploration for its consequences in lexical semantics.

1.2 Distributional models and Conceptual Reuse

“... the most interesting and important models for information retrieval, a vector space model, a probabilistic model and a logical model [...] can be described and represented in Hilbert space. The reasoning that occurs within each one of these models is formulated algebraically and ... depend essentially on the geometry of the information space.”


Distributional approaches to meaning have been successfully applied to a variety of problems related to language and beyond, thanks to their peculiar nature and properties.

A first important property is that distributional approaches allow to model different and equally rich context-sensitive semantic representations of a given language, inspired by a Wittgensteinian notion of language in use (Wittgenstein, 1953). By
varying types of context, geometry, and similarity measures, these representations are still implemented as spaces, vectors, and embedded structures, such as graphs and functions. All of these express a contextual notion of meaning that can be successfully leveraged for the use and the maintenance of the corresponding lexical knowledge.

A second relevant property of distributional approach, is that they embody a way to investigate lexical-semantic properties of a language, as a side-effect of the text mining processes that has been used to build the semantic space. This is beneficial for a truly empirical account of the language in use perspective on meaning.

Third, the set of spaces, vectors or analytical functions used as a representational device for lexical information provides a strong but unifying level of abstraction across heterogeneous cognitive representations and processes.

This third property is particularly interesting in the view of Natural Language Engineering. At the end of ’90s, Natural Language Engineering focused on the notions of algorithmic reuse as opposed to data reuse. The former were referring to the portability of individual algorithms and software tools. The latter focused on standardized annotation formalisms to favor the integration of labeled resources within different NLP processes and systems. The geometrical abstractions of distributional models support both notions. On the one side, they unify the representation formalisms, as real-valued vectors and matrices are very simple and semantically transparent data objects. On the other side, they allow a variety of algorithmic techniques, such as space transformations like projections or embeddings, to be easily ported across different tasks and linguistic levels. It can be noticed that semantic spaces enable even more sophisticated forms of reuse that we here call ‘conceptual’. A large set of heterogeneous cognitive phenomena can be in fact captured rather naturally by vector-based models (Gärdenfors, 2004).

Conceptual reuse suggests the application of geometrical paradigms as a bridge across different problems. As opposed to heterogeneous paradigms specific to sub-tasks, geometry supports higher level abstractions, as well as a stronger unifying power. An example of a successful research based on a unified paradigm is the work in (Globerson et al., 2007). The authors apply algorithms for geometrical embedding that handle different data types, such as images and texts, to search low dimensional continuous representations for semantic data. A single common Euclidean space is thus derived, based on the co-occurrence statistics of the different data types. Once the embedding is performed, “...it induces a meaningful metric between objects of the same type. Such an approach may be used, for example, for embedding images based on accompanying text, and derive the semantic distance between images.” (Globerson et al., 2007).

Several other examples of conceptual reuse can be found in the cognitive science literature. Computational approaches to cognitive models have been often advocating geometry and distributional theories as a representational paradigm able to integrate different levels of cognition (Gärdenfors, 2004; Widdows, 2004). In particular, distributional techniques have been successfully applied to studies regarding metaphor detection and analysis (Kintsch, 2000), priming (Lowe and McDonald, 2000), discourse analysis (Landauer and Dumais, 1997) and neural activation anal-
ysis (Mitchell et al., 2008), just to cite some. The underlying assumption is that the human cognitive ability of judging similarity among things and phenomena can be elegantly modeled or approximated by geometric spaces, that can provide powerful mathematical explanations of complex and effective notions of linguistic and conceptual similarity. “There is (...) no simple correspondence between cognitive functions and brain dynamics, and possibly different areas with different architecture may generate similar geometries and, hence, similar meaning content” (Fenstad, 1999).

The above approaches are just additional examples with respect to the large body of literature on distributional methods in Natural Language Engineering. Distributional models have been successfully applied to several NLP tasks, such as word clustering (F. Pereira and Lee, 1993), harvesting thesauri (Lin, 1998), word sense disambiguation (Schutze, 1998), acquisition of inference rules (Lin and Pantel, 2001) or selectional preferences (Pantel et al., 2007; Basili et al., 2007), conceptual clustering (Pantel and Lin, 2003), as well as the modeling of frame semantics phenomena (Pennacchiotti et al., 2008) or question answering (Van der Plas, 2008).

1.3 Distributional models and Natural Language Learning

The empirical nature of a distributional approach to lexical meaning enforces a strong connection with the entire field of machine learning. Contemporary research in machine learning emphasizes the role of geometrical spaces and distributional approaches in learning processes, with relevant applications to NLP tasks. Semantic spaces are relevant under this specific perspective referred as computational natural language learning. Two major specific trends are worth here to be emphasized from a distributional perspective. First, semantic spaces play a relevant role in the definition of linguistically principled kernel functions for supervised language learning. Second, they inherit a variety of algebraic techniques as extensions to basic vector space models, e.g. dimensionality reduction algorithms, able to effectively model unsupervised learning processes.

Semantic spaces and Kernels. In the mid-90ies, Statistical Learning Theory (Vapnik, 1995) shed some light on the relationships between empirical data analysis and learnability, through the mathematical characterization of the link between inductive problems and the classes of learnable functions to solve (Empirical Risk Minimization principle). This principle inspired the class of inductive algorithms known as Support Vector Machines (Vapnik, 1995). A key research area on SVMs is the study of kernel functions to model the learning problem (Haussler, 1999; Cristianini and Shawe-Taylor, 2000). Kernels efficiently increase the expressiveness of a class of functions by also preserving desirable mathematical properties such as optimality, convergence and stability. In NLP, kernels have been explored for

---

2 The author closes the paper rather radically: “And this is why grammar needs geometry more than lambda-terms.”
complex tasks, such as semantic role labeling and question answering. The so-called tree kernels (Collins and Duffy, 2002; Moschitti, 2004; Moschitti et al., 2008) explicitly model syntactic similarity.

However, kernels are nothing but scalar products in metric spaces. The semantic spaces studied by distributional approaches induce scalar products that, by definition, embody meaningful linguistic information, as exhibited by the underlying corpora. These linguistically principled metric functions may act as lexical semantic kernel functions in a variety of supervised learning algorithms. The class of semantic kernels resulting from highly dimensional word spaces constitute a very effective integrated model of different linguistic features. Distributional types of linguistic kernels usually combine syntactic, distributional and prior semantic information, as for example explored in (Bloehdorn et al., 2006; Bloehdorn and Moschitti, 2007; Ó Séaghdha and Copestake, 2008).

Algebraic transformations of Word Spaces for Unsupervised Learning.
Vector space models and linear algebra have been largely applied for feature selection and dimensionality reduction. Here, (usually linear) transformations of the original vectors are proposed to induce more expressive and efficient feature spaces. Eigenvector analysis and Principal Component Analysis give rise to a number of variants, where feature spaces obtained by pure geometrical methods are used to capture more convenient representations for the original data. Geometric transformations for dimensionality reduction have been explored since since the 90ies (Landauer and Dumais, 1997; Hofmann, 1999; Tenenbaum et al., 2000). These aim at isolate the subset of relevant information implicit in a large semantic space, and representing this information in a new space, with minimal number of dimensions. Although several applications have already shown the huge impact of these methods in terms of improved accuracy and scalability, the implications on text classification and lexical acquisition tasks have not yet been fully explored.

To date, the first most relevant work on dimensionality reduction is Latent Semantic Indexing (Berry et al., 1995), in the framework of Information Retrieval. Latent Semantic Analysis (LSA) is an algorithm first presented by (Furnas et al., 1988), and then by (Landauer and Dumais, 1997). Given an original space of documents, LSA finds the best subspace approximation of the original space, by minimizing the global reconstruction error, by projecting data along the directions of maximal variance.

More recently, new researches have emphasized the role of local information in the original space, as opposed to global ones, in those cases in which euclidean metrics have only local validity. Typical approaches try to exploit the geometry of the underlying manifold in which the data are described, in form of a variety of manifold-based learning algorithms (Tenenbaum et al., 2000; Roweis and Saul, 2000; Saul and Roweis, 2003). Isomap (Tenenbaum et al., 2000) was originally proposed as a generalization of multidimensional scaling, where symmetric adjacency graphs (based on criteria such as symmetric nearest neighborhoods) are used as opposed to Euclidean distances in the high-dimensional space. The geodesic distance on the manifold allows to unfold the original space and improve the generalization in classification and clustering. New approaches have been recently proposed,
such as Locally Linear Embedding (LLE) (Roweis and Saul, 2000) and Locality Preserving Projection (He and Niyogi, 2003). In LLE, each high dimensional input is mapped into a low dimensional output representing global internal coordinates on a locally Euclidean manifold. All these results are characterized by some geometrical assumptions, such as the clustering hypothesis, and are used to promote rich topologies out from the source data distribution. The transformations of the original representation space determines novel metrics within the original feature space.

In synthesis, this class of methods allow to learn the metrics from the data themselves, and are particularly attractive for linguistic tasks. Successful examples of these approaches are latent semantic kernels (Cristianini et al., 2002) and Manifold Regularization for semi-supervised learning (Belkin et al., 2006; Sindhwani et al., 2006).

1.4 Perspectives for distributional approaches

“We come now to the question: what is a priori certain or necessary, respectively in geometry (doctrine of space) or its foundations? Formerly we thought everything; nowadays we think nothing. Already the distance-concept is logically arbitrary; there […]”

(A. Einstein)

Metric spaces are used in many natural sciences, such as biology and physics, to model domains, as typed feature structures in (Pollard and Sag, 1994). We can also see geometry, i.e. analytical/descriptive as well as differential geometry, as the modeling domain for language semantics. Reasoning over mathematical abstractions such as vectors and probability (i.e. density) functions can be reasonably expected to say a lot about intrinsic properties of natural language semantics. This perspective is nowadays even more critical, as for the urgency of understanding the interplays between the social and linguistic dimensions of text semantics in the Web 2.0 era. Geometrical models are here useful along two directions.

First, semantic spaces provide a framework where linguistic phenomena can be studied through a mathematics that helps to account for the discrete but non linear nature of linguistic phenomena (Manifold Learning). The manifold assumption states that a usually low-dimensional possibly non linear manifold accounts for the distributions of observable data, and that categorical knowledge can be seen as a smooth function over these manifolds. The application of this perspective in NLE studies consists in the search of linguistically motivated manifolds that correspond to useful generalizations. As a consequence, weakly supervised (or even unsupervised) learning models over these manifolds can be designed. A large family of semi-supervised learning methods have been recently proposed to exploit the distributions of unlabeled data, as a way to maximize the generalization accuracy by learning from small sets of labeled data (Sindhwani et al., 2006). In such a way, unlabeled examples are used to derive information about the underlying implicit manifold, that is very likely non Euclidean. They maximize the predictive
accuracy of known supervised algorithms (e.g. SVMs) by exploiting the geometry of the underlying manifold, i.e. the distributions intrinsic to the linguistic data sample. As the amount of unlabeled data is usually very large in language learning problems, this view is very attractive in NLE although not yet fully explored. Moreover, this perspective is even more appealing as the information about the manifold can in principle be also gathered from a priori linguistic knowledge. In facts, nothing prevents to impose a priori restrictions on the unlabeled data, so that their distribution is forced to the specific interpretation inspired by a resource, e.g. a semantic dictionary. In this case, the manifold is artificially superimposed, but the mathematics of the above manifold-based semi-supervised learning is fully preserved. The result is an elegant way of integrating lexical knowledge as a set of a priori constraints over an embedding into the manifold in which the learning takes place. A noticeable example applied to sentiment analysis is recently discussed in (Sindhwani and Melville., 2008). Further work where dimensionality reduction methods are applied to elegantly integrate a priori information is (Weinberger and Chapelle, 2008), where a topic taxonomy is embedded within a document representation space, and accordingly an effective semi-supervised document classification method is obtained.

Vector Spaces are also important models for analytical descriptions of uncertain systems, and in particular incomplete systems. The physiological and psychological aspects of language have been often recognized as out of the scope of systems characterized by a complete knowledge. In analogy with quantum physics, incompleteness is rather standard for the knowledge available to many linguistic inferences. The logic developed in this frameworks, i.e. quantum logic (Birkhoff and von Neumann, 1936), is a model of inference with respect of a variety of uncertain conditions. The crucial properties for such logics are, among others, graded notions of similarity and non commutativity of several operators (Varadarajan, 1985). In (van Rijsbergen, 2004) well known measures, as projections in subspaces, are discussed as significant non commutative operators. While measurement in classical mechanics is always commutative, quantum mechanics allows naturally for its non commutativity. Composition of projections as well as tensor products, are both useful operators for directional inference in quantum mechanics. The impact of these possibilities on language processing models is clear. Language technologies can look at semantic spaces with a strong interest not only on symmetrical relations (e.g. as the notion semantic similarity measures traditionally modeled for sense discrimination or disambiguation tasks, (Schutze, 1998)), but mainly for complex directional linguistic inferences, e.g. textual entailment.

2 Overview of this volume

The following list includes all the papers accepted in this special issue that tackle a broad spectrum of problems related to distributional models of lexical semantic phenomena:
Distributional Lexical Semantics

- Directional Distributional Similarity by Lili Kotlerman, Ido Dagan, Idan Szpektor and Maayan Zhitomirsky-Geffet
- A Non-negative Tensor Factorization Model for Selectional Preference Induction by Tim Van de Cruys
- Inductive Probabilistic Taxonomy Learning using Singular Value Decomposition by Francesca Fallucchi and Fabio Massimo Zanzotto
- A Class-based Approach to Disambiguating Levin Verbs by Jianguo Li and Chris Brew
- Automatic Discovery of Word Semantic Relations using Paraphrase Alignment and Distributional Lexical Semantics Analysis by Gal Dias, Rumen Moraliyski, Joo Cordeiro, Antoine Doucet and Helena Ahonen-Myka
- Modeling Reciprocal Social Interactions with Latent Space Models by Roxana Girju and Michael Paul
- An information-theoretic, vector-space-model approach to cross-language information retrieval by Peter Chew, Brett Bader, Stephen Helmreich, Ahmed Abdelali and Stephen Verzi
- The Automatic Identification of Lexical Variation between Language Varieties by Yves Peirsman, Dirk Geeraerts, Dirk Speelman

The first three papers focus on general computational aspects. The adoption of vector spaces as representational paradigms gives rise to here novel algorithmic techniques able to improve the acquisition processes or the operational inferences regarding textual or lexical semantic phenomena. Lexicon-oriented papers form a second group that focuses on distributional methods as tools for the acquisition of specific components of the lexicon – e.g. the acquisition of selectional preferences for large scale semantic lexicons. Finally, a third group of papers clusters works on novel NLP applications supported by more or less complex distributional models – e.g. the recognition and modeling of social relationships in texts from open sources.

The general computational aspects group includes three papers: the work by Kotlerman and colleagues on directional models of text similarity, the paper by Van de Cruys on a tensor analysis method, and the work by Fallucchi and Zanzotto on the use of SVD-based techniques for efficient probability estimation in a taxonomy learning task. These three works extend in an original way the traditional repertoires of distributional analysis tools.

The second group of lexicon-oriented papers is characterized by three works devoted to the automatic acquisition of lexical information, these including the mentioned work by Van de Cruys; the work by Li and Brew on the disambiguation of verb semantic classes, and the work by Gaël Dias and colleagues, on the discovery of semantic relations among words.

Finally, the NLP applications group includes three papers. Girju and Paul model social interactions over Social Web data based on a distributional account of specific linguistic patterns. Chew and colleagues apply an information theoretic weighting scheme in a vector space model, for cross-language Information Retrieval. Finally, Peirsman and colleagues, present an original application of distributional analysis on parallel corpora, as a tool for studying variational linguistics.
The large number of contributions could not be allocated on this volume for editorial limitations. Three papers, i.e. the paper by Fallucchi and Zanzotto, from Girju and Paul, and finally the paper from Chew and colleagues, have been moved to the next following issue. This choice is only tailored to a better fitting of the space limitations of individual journal volumes, and it is independent from any technical quality issue or criteria.

The overall special issue sheds more light on a huge research area, which is also testified by an increasing number of conferences and workshops dedicated to distributional semantics. As an example, the 2010 edition of the ACL Conference includes three workshops highly related to distributional lexical semantic topics, that is “TextGraphs-5: Graph-based Methods for Natural Language Processing”, “GEMS 2010: Geometrical Models of Natural Language Semantics” and “Domain Adaptation for Natural Language Processing (DANLP)”.

We would like to thank the Editorial Board members of this Special Issue, for their invaluable support. Their hard work will allow these topics to be better understood and fruitfully applied in future research.

References


