Research Statement for Irina Matveeva

Department of Computer Science
University of Chicago
Chicago, IL 60637
matveeva@cs.uchicago.edu

My research interest is to improve natural language applications by developing efficient unsupervised and semi-supervised machine learning approaches. My approach is to design machine learning solutions tailored to specific natural language problems based on an in-depth analysis of their components. I believe that machine learning algorithms are most efficient for language applications if they have an elegant linguistic interpretation.

A better understanding of document content is one of the main goals of Natural Language Processing (NLP). As natural language applications not only retrieve information for the user but also provide an increasingly detailed analysis of it, they need linguistic knowledge and the ability to analyze semantic relations. A key challenge of performing large-scale semantic analysis is to avoid using manually built linguistic resources which are expensive to construct. The main focus of my research is on document indexing and representation of term-document relations as the basic steps in analyzing document's content. I generalized the idea of Latent Semantic Analysis (LSA) and developed a flexible unsupervised framework for extracting semantic relations with dimensionality reduction methods which improved the performance of several applications [9, 8, 7].

More broadly, I am interested in developing notions of similarity for different types of linguistic data. I successfully employed dimensionality reduction and graph-based machine learning in diverse applications ranging from synonymy induction and bilingual dictionary extraction to information retrieval, document classification, and automated morphology learning [4, 3, 2, 6, 5].

Generalized Latent Semantic Analysis A multidimensional vector space representation for documents is required as input in many state-of-the-art approaches. It is convenient and makes similarity measures such as cosine similarity or distance immediately available. However, these measures will not be effective if they do not have a natural interpretation for the original text data.

Latent Semantic Analysis (LSA) is the most influential dimensionality reduction algorithm in NLP which has a linguistic interpretation. LSA puts documents in the space of latent semantic concepts broadly corresponding to the topics in the document collection. However, LSA is a special case in a general approach of multidimensional scaling and is restricted to dual term-document representation which negatively impacts its performance on heterogeneous document collections.

I developed the Generalized Latent Semantic Analysis (GLSA) framework to extend the applicability of the idea of the latent semantic space analysis [9]. GLSA computes a semantically motivated term and document representation and improves the performance on the TOEFL synonymy test, and on document clustering and classification [9, 8, 7]. GLSA is particularly useful for applications that work with sentences and short paragraphs. Such applications often require a fine-grained analysis of semantic similarity between words because they do not have much disambiguating context. GLSA improves the quality of automated multi-document summaries, especially for heterogeneous documents that don't have much word overlap.

GLSA brings together the ideas from unsupervised similarity induction and dimensionality

reduction and computes a word and document representation which preserves the true semantic similarities [9]. The main advantage of GLSA is its flexibility. It uses pair-wise semantic similarities between words which can be computed in different ways, in particular with corpus-based co-occurrence statistics. Furthermore, GLSA employs different methods of dimensionality reduction to preserve the original similarities. Currently, I am working on the theoretical analysis of the combination of GLSA with graph-based dimensionality reduction and with language modelling.

Graph-based GLSA The notion of semantic similarity between words is intuitive and it is meaningful to define a linear order relation between semantically close words, whereas it is more difficult to define the notion of semantic dissimilarity. The distinction between similar and dissimilar words can be captured effectively by using a similarity graph. Documents and words can be naturally represented as nodes in a similarity graph. Two nodes are connected in the graph if the underlying objects are considered similar. I am extending my approach with graph-based GLSA to restrict it to modelling the similarities between words that are closely related [8].

Language Modelling I am interested in combining the idea of GLSA with probabilistic generative approaches. Probabilistic methods are successfully used to model the relations between words and word-document relations. However, the parameter estimation for these models can be computationally demanding with the increased complexity of the model. I proposed a simple and effective method of employing GLSA-based similarities in an extended language model which accounts for synonymy and polysemy [7]. Currently, I integrate random walks and spectral analysis of the word similarity graph to model second and higher-order semantic relations between words and documents.

Unsupervised Morphology Induction Unsupervised learning of grammar is another level of using corpus-based similarity and graph-based methods to improve linguistic analysis without manually built resources. As part of the *Linguistica* group, I worked on unsupervised morphology learning. I contributed to the improved string edit distance heuristic for learning languages with rich morphology [3]. I also worked on inducing morphological classes of words as bootstrapping for unsupervised learning of grammatical categories. In this project we started with the automated morphology learner that produces fine-grained morphological classes and clustered them using syntactic context and graph partitioning. I showed that syntactic similarity is a good means of building coarser classes which reveal the connection between morphology and lexical categories and thus reduce the description length of the model [4].

Semi-supervised Parsing One of the recent exciting areas of application for machine learning in NLP are problems with structured output. I am using transductive and semi-supervised approaches for these applications because unsupervised learning in this case would not be accurate enough. Parsing is one of the particularly challenging applications which typically requires a large number of manually labeled data.

I am currently working on transductive parsing learning which reduces the need for manual labels by exploiting the similarity structure of the data. My approach is to construct a similarities graph for parse trees and use the graph transduction algorithm [1] to rank them. The advantage of graph transduction is that it does not require any representation for parse trees as input. One of the main challenges in this setting is to define a notion of similarity for complex structures such as a parse tree in order to compute a good similarity graph. I explore different notions of similarity

with emphasis on the hierarchical structure of parse trees and study the similarity structure of the space of the parse trees to learn the notion of a good parse.

Future Work So far, I have considered the notion of similarity in a uniform manner. My work on parsing shows, however, that for objects with hierarchical structure it is more appropriate to have a separate definition of similarity for different levels. I believe that a similar approach can be used for text data since semantics and syntax provide a theoretical justification for treating subsets of the vocabulary differently. For instance, two documents can be similar either because they contain information about the same named entities or because they contain semantically related words.

I believe that the multi-level model of document similarity is particularly useful for the information available on the Web because it combines different writing styles and media types. I am going to extend my work to NLP applications, such as question answering and lexical entailment, that need the ability to capture the semantic equivalence between different pieces of information and therefore require particularly deep knowledge of syntax and semantics. It will be exciting to adapt these applications to new challenges such as integration of text and image analysis and informal writing styles of blogs and other sources of user generated content. My main approach will continue to be the combination of unsupervised and semi-supervised machine learning solutions with linguistic modelling to tackle the challenges of robust large scale automated text understanding.

References

[1] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Proceedings of Conference on Learning Theory (COLT)*, pages 624–638, 2004.

[2] E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Djean. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 526–533, 2004.

[3] Y. Hu, I. Matveeva, J. Goldsmith, and C. Sprague. Refining the string edit distance heuristic for morpheme discovery: Another look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition (PsychoCompLA) at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 28–36, 2005.

[4] Y. Hu, I. Matveeva, J. Goldsmith, and C. Sprague. Using morphology and syntax together in unsupervised learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition (PsychoCompLA) at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 20–28, 2005.

[5] G. Levow and I. Matveeva. University of Chicago at Cross-language Evaluation Forum 2004: Coss-language text and spoken document retrieval. In *Proceedings of the Workshop of the Cross-Language Evaluation Forum (CLEF), Lecture Notes in Computer Science*, volume 3491, pages 170–179, 2004.

[6] I. Matveeva, C. Burges, and T. Burkard. High accuracy retrieval with Multiple Nested Ranker. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 437 – 444, 2006.

[7] I. Matveeva and G.-A. Levow. Computing term translation probabilities with Generalized Latent Semantic Analysis. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL) (poster presentation)*, 2006.

[8] I. Matveeva and G.-A. Levow. Graph-based Generalized Latent Semantic Analysis for document representation. In *Proceedings of the TextGraphs Workshop at the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pages 61–64, 2006.

[9] I. Matveeva, G.-A. Levow, A. Farahat, and C. Royer. Generalized Latent Semantic Analysis for term representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 60–68, 2005.