Radu Soricut

Contact Information	1440 12th St. Unit C Manhattan Beach CA 90266 USA	*Voice:* (310) 754-5968 *E-mail:* radu.soricut@gmail.com *Web:* www.radusoricut.com

Research Interests

Natural Language Processing: statistical machine translation, language generation, language modeling, syntactic parsing, discourse parsing, question answering, automatic summarization

Machine Learning: supervised and unsupervised learning, quality prediction for structured output

Education

Ph.D., Computer Science

2006, University of Southern California, Los Angeles, California USA

M.Sc., Computer Science

2003, University of Southern California, Los Angeles, California USA

2001, University of Iowa, Iowa City, Iowa USA

B.Sc., Computer Science

1997, University of Bucharest, Bucharest, Romania

Employment

SDL International, Language Weaver, Los Angeles, California USA

Senior Research Scientist, Manager of Applied Sciences and Engineering **August 2010 - present**

Leading an R&D team to design and implement a production system for building domain-specialized translation engines. Determining the methodology and the processes via which such systems are built. Delivering domain-specialized translation engines at performance levels that are consistently higher compared to generic, non-specialized translation engines.

Leading an R&D team to create and develop a wide-range of prediction capabilities for automated translation output. A quality-prediction system built by this team won 1st place in the WMT2012 shared-task for Machine-Translation Quality Estimation (out of 20 participating submissions).

Language Weaver, Inc., Los Angeles, California USA

Research Scientist **September 2006 - July 2010**

Designed and developed TrustRank and TrustScore, algorithms that use Machine Learning techniques to learn to predict a numerical score that a human would likely assign for the quality of an automatically-translated document. These components allowed Language Weaver, Inc. to create a go-to-market strategy around the concept of "trusted translations". Language Weaver won a 2009 Innovator Award from Global Service and Support Associations for the TrustScore capability.

Designed and developed InvestmentPredictor, a capability for assessing the level of investment (in terms of training data quantity and dollar figures) needed for translation systems to achieve certain accuracy levels. It also offers the capability of predicting the performance level of yet-to-be-built translation systems with predefined configurations.

Designed and implemented a data selection algorithm (using active-learning techniques) that maximizes the impact of data on system quality performance at a given cost. The use of this algorithm was assessed to save between 30% and 70% of the costs associated with producing training data for statistical translation systems.

Designed and implemented algorithmic improvements to statistical translation algorithms, using syntactic parse trees to inform statistical translation decisions.

Designed a quality predictor for syntactic parsing that allows decision-making in terms of parsing accuracy on various domains.

Multiple patent applications filed regarding these contributions.

Received the Language Weaver Superstar Award in 2009 for these contributions.

Information Sciences Institute, University of Southern California, Los Angeles, USA

Ph.D. Candidate/Research Assistant **2001-2006**

Ph.D. thesis: Natural Language Generation using an Information-Slim Representation. This work describes a formal representation that can be used to support natural language applications both on the language understanding and the language generation side. It demonstrates the advantages of this formal representation for applications such as automatic translation, automatic summarization and document generation.

Designed and implemented a statistical discourse parser that analyzes its input text to create hierarchical representations of discourse structure. This work has applications in natural language applications such as automatic summarization, automatic essay scoring, and dialog generation.

Microsoft Research, Redmond, Washington USA

Research Intern **Summer 2003**

Designed and developed an automatic question-answering engine. The algorithm uses unsupervised learning techniques to bridge the gap between question and answer lexical items from FAQ pages. The question-answering engine is capable of finding non-factoid answers using an off-the-shelf search engine.

Designed a generic framework for evaluating natural language systems on a variety of tasks. It accommodates evaluations for automatic summarization, automatic translation, and automatic question-answering.

Institute for Information Sciences and Technologies, Macau

Research Fellow **1996-1997**

Used a formal algebraic framework to prove properties related to telecommunication protocol specifications. This work had direct applications in proving correctness and specification compliance in the telecommunication industry.

Honors and Awards

Company Superstar Award, Language Weaver, Inc., 2009

Microsoft Research Fellowship Finalist, 2005

Gerard P. Weeg Scholarship, University of Iowa, 2000

Rockwell Collins Scholarship, University of Iowa, 1999

Research Fellowship, United Nations University, Institute for Information Sciences and Technology, 1996

Undergraduate Fellowship of Merit, University of Bucharest, 1992-1997

Patents Awarded

Radu Soricut and Daniel Marcu. 2011 (Awarded), 2006 (Submitted). Weighted System of Expressing Language Information Using a Compact Notation. US Patent 7,974,833. July 2011.

Radu Soricut, Daniel Marcu and Kevin Knight. 2008 (Awarded), 2002 (Submitted). Statistical Translation Using a Large Monolingual Corpus. US Patent 7,430,388. March 2008.

Patents Pending

Radu Soricut and Daniel Marcu. 2012 (Submitted). Trust Scoring for Language Translation Systems. Submitted to the United States Patent and Trademark Office. June 2012.

Radu Soricut, Swamy Viswanathan, and Daniel Marcu. 2010 (Submitted). Multiple Means of Trusted Translation. Submitted to the United States Patent and Trademark Office, Application 12/820,061 (Continuation-in-Part to 12/572,021). June 2010.

Radu Soricut, Swamy Viswanathan, and Daniel Marcu. 2010 (Submitted). Predicting the Cost Associated with Translating Textual Content. Submitted to the United States Patent and Trademark Office. March 2010.

Radu Soricut, Swamy Viswanathan, and Daniel Marcu. 2009 (Submitted). Providing Machine-Generated Translations and Corresponding Trust Levels. Submitted to the United States Patent and Trademark Office, Application 12/572,021. September 2009.

Daniel Marcu, Radu Soricut, and Swamy Viswanathan. 2009 (Submitted). Translating Documents Based on Content. Submitted to the United States Patent and Trademark Office. July 2009.

Peer-Reviewed Publications

Radu Soricut and Sushant Narsale. 2012. Combining Quality Prediction and System Selection for Improved Automatic Translation Output. Proceedings of the ACL Seventh Workshop on Statistical Machine Translation (WMT-2012). June 7-8, Montreal, Canada.

Radu Soricut and Nguyen Bach and Ziyuan Wang. 2012. The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task. Proceedings of the ACL Seventh Workshop on Statistical Machine Translation (WMT-2012). June 7-8, Montreal, Canada.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. Proceedings of the ACL Seventh Workshop on Statistical Machine Translation, (WMT-2012). June 7-8, Montreal, Canada.

Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. Proceedings of the Association for Computational Linguistics Conference (ACL-2010). July 11-16, Uppsala, Sweden.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic Prediction of Parser Accuracy. Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP-2008). October 25-27, Waikiki, Honolulu, Hawaii.

Radu Soricut and Daniel Marcu. 2007. Abstractive headline generation using WIDL-expressions. Journal of Information Processing and Management, 2007, 43(6):1536-1548.

Radu Soricut. 2006. Natural Language Generation using an Information-Slim Representation. Ph.D. Thesis, University of Southern California, Department of Computer Science, 2006.

Radu Soricut and Eric Brill. 2006. Automatic Question Answering Using the Web: Beyond the Factoid. Journal of Information Retrieval - Special Issue on Web Information Retrieval, 2006, 9:191-206.

Radu Soricut and Daniel Marcu. 2006. Stochastic Language Generation Using WIDL-expressions and its Application in Machine Translation and Summarization. Proceedings of the Association for Computational Linguistics Conference (ACL-2006). July 17-21, Sidney, Australia.

Radu Soricut and Daniel Marcu. 2006. Discourse Generation Using Utility-Trained Coherence Models. Proceedings of the Association for Computational Linguistics Conference (ACL-2006). July 17-21, Sidney, Australia.

Radu Soricut. 2005. Natural Language Generation for Text-to-Text Applications Using an Information-Slim Representation. Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-2005), AAAI/SIGART Doctoral Consortium. July 9-13, Pittsburgh, PA.

Radu Soricut and Daniel Marcu. 2005. Towards Developing Generation Algorithms for Text-to-Text Applications. Proceedings of the Association for Computational Linguistics Conference (ACL-2005). June 25-30, Ann Arbor, MI.

Radu Soricut and Eric Brill. 2004. A Unified Framework for Automatic Evaluation using N-gram Co-Occurrence Statistics. Proceedings of the Association for Computational Linguistics Conference (ACL-2004). July 22-25, Barcelona, Spain.

Radu Soricut and Eric Brill. 2004. Automatic Question Answering: Beyond the Factoid. Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-2004). May 2-5, Boston, MA.

Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL-2003). May 27-June 1, Edmonton, Canada.

Radu Soricut, Kevin Knight, and Daniel Marcu. 2002. Using a large monolingual corpus to improve translation accuracy. Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-2002). Tiburon, CA, October 8-12.

Hal Daume III, Abdessamad Echihabi, Daniel Marcu, Dragos Stefan Munteanu, and Radu Soricut. 2002. GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries. Proceedings of the Document Understanding Conference (DUC-2002). Philadelphia, PA, July 11-12.

Professional Service

Workshop Co-Organizer: ACL Seventh Workshop on Statistical Machine Translation (WMT-2012), Quality-Estimation Track

Journal Reviewing: Computational Linguistics, Journal of Artificial Intelligence Research, Dialog and Discourse

Conference Reviewing: ACL, NAACL, HLT, EMNLP

Technical Skills

Natural Language Processing: machine translation, language modeling, syntactic parsing, discourse parsing, automatic question answering, automatic summarization, evaluation metrics for NLP tasks

Machine Learning: supervised learning (regression, classification), active learning, semi-supervised learning (self/co-training), unsupervised learning, minimum-error training

Programming Languages: C/C++, Perl, Python, Unix shell scripts, Pig (MapReduce framework).

Operating Systems: Unix/Linux, Windows.