Thamar Solorio

2. Education

Ph.D. in Computer Science September 2005
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México
Dissertation title: *"Improvement of Named Entity Tagging by Machine Learning"*
Advisor: Prof. Aurelio López López

M.S. in Computer Science August 2002
Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México
Thesis title: *"Using Unlabeled Data to Improve Classifier Accuracy"*
Advisor: Prof. Olac Fuentes

B.S. in Computer Systems Engineering July 2000
Facultad de Ingeniería, Universidad Autónoma de Chihuahua, Chihuahua, México

3. Professional Experience

Visiting NLP Scientist September 2021 -
Bloomberg LP

Professor September 2021 -
Department of Computer Science, University of Houston

Associate Professor September 2014 – August 2021
Department of Computer Science, University of Houston

Assistant Professor June 2009 –August 2014
Department of Computer and Information Sciences, University of Alabama at Birmingham

Research Associate September 2007–June 2009
Department of Computer Science, University of Texas at Dallas

Lecturer September 2005–August 2007
Department of Computer Science, University of Texas at El Paso

4. Research Interests

Natural Language Processing and Applied Machine Learning: natural language processing technology for mixed-language data, stylistic modelling of text, information extraction from social media data, multimodal processing of videos

5. Honors, Recognitions and Outstanding Achievements

UHCS Academic Excellence Award Fall 2017
COSC Department Award

CRA-W Mid Career Mentoring Workshop June 2015
Travel Grant

2014 Denice Denton Emerging Leader ABIE Award Fall 2014
Recognition by the Anita Borg Institute

Research Visit for Distinguished Researchers, UPV, Spain Fall 2011
Research Visit Grant

Anthony Barnard Award, CIS, UAB Fall 2011
Travel Grant

Anthony Barnard Award, CIS, UAB Fall 2009
Travel Grant

CRA Academic Careers Workshop, NSF February 2008
Travel Grant

CRA-W Career Mentoring Workshop, NSF June 2007
Travel Grant

National Council for Science and Technology (CONACyT) September 2002- August 2005
Fellowship for PhD studies

National Council for Science and Technology (CONACyT) August 2000- August 2002
Fellowship for MS studies

6. Peer-Reviewed Competitive Research Grants Funded (over $3M in funds secured)

OISE, Office Of International Science & Engineering, Div. of Information & Intelligent Systems, NSF September 2021-August 2024
$299,716.00 Track I: US-Mexico Collaboration on Multimodal Detection of Objectionable Content in Online Videos in Spanish and English
Investigators: Thamar Solorio (PI, UH), Ioannis Kakadiaris (co-PI, UH)

IIS, Div. of Information & Intelligent Systems, NSF August 2020-July 2021
$43,959 Workshops on desiderata for a multimodal dataset for objectionable content detection
Investigators: Thamar Solorio (PI, UH), Ioannis Kakadiaris (co-PI, UH)

IIS, Div. of Information & Intelligent Systems, NSF May 2020-May 2021
$16,000 REU Supplement
Investigators: Thamar Solorio (PI, UH)

IIS, Div. of Information & Intelligent Systems, NSF October 2019-August 2022
$307,918 RI: Small: Robust Models for Sequence Labelling in Social Media Data
Investigators: Thamar Solorio (PI, UH)

Div. of Industrial Innovation and Partnerships, NSF July 2018–July 2019
$225,000 SBIR Phase I: VideoPoints: A Companion for Classroom Learning
Investigators: Tayfun, Tuna (PI,VideoPoints), Jaspal Subhlok, (co-PI, VideoPoints), Shishir Sha (co-PI, VideoPoints), Thamar Solorio (co-PI, VideoPoints)

U.S. Department of Defense July 2016–July 2017
$406,963 Cluster Computing Infrastructure for Large Scale Data and Text Analytics at UH
Investigators: Rakesh Verma (PI, UH) Thamar Solorio (co-PI, UH)

IIP, STTR Phase I, NSF January 2016–December 2016
$224,110 STTR Phase I: Book Discovery through Literary DNA

Holly Payne (PI, Booxby) Thamar Solorio (co-PI, UH)

IIS, Secure & Trustworthy Cyberspace, NSF August 2015–July 2018
$499,695 TWC: Small: Statistical Models for Opinion Spam Detection Leveraging Linguistic and Behavioral Cues
Arjun Mukherjee (PI, UH) Thamar Solorio (Senior Personnel, UH)

IIS, Robust Intelligence, NSF January 2014–January 2019
$470,000 CAREER: Authorship Analysis in Cross-Domain Settings
Thamar Solorio (PI, UH)

CNS, Div. of Computing and Network Systems, NSF September 2012–August 2015
$364,301, CI-ADDO-NEW: Collaborative Research: A Repository for Annotating Multilingual Code Switched Data
Investigators: Thamar Solorio (PI, UH)

Human-Centered Computing, IIS, NSF September 2010–August 2016
$301,055, Analysis of Language Impairment in Monolingual and Bilingual Children
Thamar Solorio (PI, UH)

IIS, Div. of Information & Intelligent Systems NSF September 2012–August 2014
$45,048, EAGER: Investigating linguistic dimensions in cross-domain authorship analysis
Thamar Solorio (PI, UAB)

Office of Naval Research March 2012–February 2015
$379,171, Secure Document Signatures
Thamar Solorio (PI, UAB)
Ragib Hasan (co-PI, UAB)

Robust Intelligence, IIS, NSF January 2011–December 2011
$16,200, ACL-HLT Student Session
Investigators: Thamar Solorio (PI, UAB)

Computing Research Infrastructure, NSF February 2010–February 2012
$21,992, Collaborative Research:CI-P: Creation of an annotated repository of multilingual and multigenre code switched data for several language pairs
Investigators: Thamar Solorio (PI, UAB)

Robust Intelligence and OISE, NSF January 2010–July 2010
$17,972, Young Investigators in the Americas Workshop
Investigators: Thamar Solorio (PI, UAB)

Human-Centered Computing, IIS, NSF September 2008–August 2010
$110,493, HCC-Small:Collaborative Research: Prediction of Language Status in Bilingual Children (PoLSi-BC)
Thamar Solorio (PI, UAB), Yang Liu (co-PI, UTD)

National Center for Research Resources, NIH September 2008–August 2009
$36,000, Prediction of Language Status in Monolingual Children (PoLSi-MC)
Thamar Solorio (PI, UTD), Yang Liu (co-PI, UTD)

7. Industry Gifts

Adobe Inc. August 2018
Unrestricted Gift Grant, $10,000

Adobe Inc. November 2018
Unrestricted Gift Grant, $10,000

Adobe Inc. February 2019
Unrestricted Gift Grant, $5,000

Adobe Inc. May 2019
Unrestricted Gift Grant, $5,000

Adobe Inc. August 2019
Unrestricted Gift Grant, $10,000

Adobe Inc. May 2020
Unrestricted Gift Grant, $7,000

Adobe Inc. September 2020
Unrestricted Gift Grant, $8,000

Adobe Inc. May 2020
Unrestricted Gift Grant, $7,000

Adobe Inc. June 2021
Unrestricted Gift Grant, $8,000

Adobe Inc. September 2021
Unrestricted Gift Grant, $8,000

Adobe Inc. March 2022
Unrestricted Gift Grant, $7,000

Adobe Inc. May 2022
Unrestricted Gift Grant, $15,000

8. Patents

Book Analysis Recommendation, Co inventors: Holly Payne, Mark Bergman, Bogart Vargas, **Thamar Solorio**, Suraj Maharjan[s] and Sudipta Kar[s]. US Patent App. 16/678,553

9. Keynote Talks

Achieving Human-Level Multilinguality. The Next Big Ideas Panel, ACL 2022.

Moving the needle in NLP technology for the processing of code-switched language, NAACL 2021.

Recent Findings on Multimodal Prediction Systems, Keynote Speaker, XXXVI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2020).

Enabling Technology for Code-Switching Data, Keynote Speaker, Linguistic Symposium on Romance Linguistics 50 [LSRL50], July 2020.

Prediction of Movie Ratings with a Multimodal Approach, Keynote Speaker, 2020 Mexican Summer School in NLP, September 2020.

Named Entity Recognition in Challenging Contexts, Keynote Speaker at the 16th Mexican Conference on Artificial Intelligence, Ensenada, Baja California, Mexico, October 2017.

Mixed-language data in social media, Keynote speaker at the Social NLP workshop, Held in conjunction with EMNLP 2017, Valencia, Spain.

No cambies de Tema, de todas formas podemos identificarte (*Don't change the subject, we can still identify you*), Keynote speaker at WOPATEC 2014, Workshop de Procesamiento Automatizado de Textos y Corpus, Viña del Mar, Chile, November 2014.

10. Other Invited Talks and Tutorials

NLP-enabled Design Assistance for Visual Communication. Invited Talk, LatinX in AI EMNLP 2020.

NLP-enabled Design Assistance for Visual Communication. Invited Talk, WeCNLP 2020.

Summer School on Natural Language Processing and Text Mining, Universidad Nacional de Colombia, Bogotá, Colombia, June 2016.

Code-Switched data: A special type of multilingual data and the challenges for NLP. Tutorial at the First Fall School in Natural Language Processing, Puebla, Mexico, November 2015.

Exploiting similarities in writing styles to predict authorship, Computer Science Colloquium, The University of Texas at Dallas, March 2013.

Authorship Analysis: Exploring Similarities Across Different Linguistic Dimensions, The Theodore Haddin Arts and Sciences Forum, UAB. January 2013.

Natural Language Processing for Computer Forensics. Invited talk at the 8th Workshop on Human Language Technologies, held in conjunction with MICAI 2011, Puebla, Mexico. November 2011.

Analysis of Language Samples for Assessing Language Status in Monolingual and Bilingual Children, University of Alicante, Alicante, Spain. September 2011.

Generating Informative Features for Authorship Identification on Social Media, Polytechnic University of Valencia, Valencia, Spain. September 2011.

Exploring New Frontiers in Natural Language Processing (NLP): Applications of NLP to Communication Disorders and Bilingual Discourse. Research Group in Computational Linguistics, University of Wolverhampton. May 2009.

Exploring corpus-based approaches for distinguishing children with and without language impairment. Callier Center, The University of Texas at Dallas. November 2008.

The Automated Processing of Bilingual Discourse. Language and Information Technologies, University of North Texas. July 2008.

Processing Code-Switched Text. Department of Computer Sciences, The University of Texas at Austin. May 2008.

Processing Code-Switched Text. Department of of Computer and Information Sciences, University of Delaware. February 2008.

11. Selected Service to the Profession/Academic Discipline

Program Co-Chair

2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-2019.

Workshop Co-Chair

Fifth Workshop on Computational Approaches to Linguistic Code Switching, collocated with NAACL 2021.
Fourth Workshop on Computational Approaches to Linguistic Code Switching, collocated with LREC 2020.
Second Workshop on Stylistic Variation, to be held in conjunction with NAACL 2018.
Third Workshop on Computational Approaches to Linguistic Code Switching, to be held in conjunction with ACL 2018.
First Workshop on Stylistic Variation, held in conjunction with EMNLP 2017.
Second Workshop on Computational Approaches to Code Switching, held in conjunction with EMNLP 2016.
First Workshop on Computational Approaches to Code Switching, held in conjunction with EMNLP 2014.
NAACL-HLT 2010, Young Investigators Workshop on Computational Approaches to Languages of the Americas

Steering Committees

First Emerging Leaders Workshop. Faculty development workshop sponsored by the Anita Borg Institute and the National Science Foundation, June 3, 2016 in Madison, Wisconsin.

Editorial Board Member

Co-Editor in Chief, ACL Rolling Review
Northern European Journal of Language Technology (NEJLT)
Computer Speech and Language

Elected Board Member

North American Chapter of the Association for Computational Linguistics, 2021, 2022.

Program & Organization Committees

ACL 2022, Senior Area Chair
SEM 2021, Area Chair, Sentiment analysis and argument mining
ACL-IJCNLP 2021, Area Chair, Computational Social Science and Social Media
NAACL 2021, Area Chair, Sentiment Analysis and Stylistic Analysis
EACL 2020, Area Chair, Sentiment Analysis and Argumentation
EMNLP 2020, Senior Area Chair, NLP Applications

WiNLP 2020, Widening NLP Workshop 2020, program committee
LREC 2020, Language Resources and Evaluation Conference
FLAIRS 2020, The 33rd International FLAIRS Conference
SEPLN 2020, 36th Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
AAAI 2019, Senior Program Committee
SEPLN 2019, XXXV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
ACL 2018 Demonstration Papers, Co-chair
SEPLN 2018, XXXIV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
ACL 2017, Program Committee
EMNLP 2017, Area Chair, Text Mining and NLP Applications
Microsoft Research India Summer Workshop on Artificial Social Intelligence, 2017
RANLP 2017, Program committee member
ICON 2017, The Fourteenth International Conference on Natural Language Processing
SEPLN 2017, XXXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
IJCNLP 2017, Area Chair, Sentiment Analysis and Opinion Mining
EMNLP 2016, Conference on Empirical Methods in Natural Language Processing
IBERAMIA 2016, 15th edition of the Ibero-American Conference on Artificial Intelligence, NLP Area Chair
ACL 2016, The 54th Annual Meeting of the Association for Computational Linguistics, area co-chair for multilinguality track
NAACL 2016, 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies
AAAI 2016, Thirtieth AAAI Conference on Artificial Intelligence
WWWW 2016, 25th International World Wide Web Conference, Natural Language Processing PC member
ACL 2015, The 53rd Annual Meeting of the Association for Computational Linguistics
SEPLN 2015, XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural
NAACL-HLT, the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), Tutorial co-chair
IBERAMIA 2014, 14th edition of the Ibero-American Conference on Artificial Intelligence
EACL 2014, 14th Conference of the European Chapter of the Association for Computational Linguistics
ACL 2014, The 52nd Annual Meeting of the Association for Computational Linguistics
RANLP 2013 Student Research Workshop
IJCAI 2013, 23^{rd} International Joint Conference on Artificial Intelligence
ACL 2013, Area Chair
CICLing-2013, 14^{th} International Conference on Intelligent Text Processing and Computational Linguistics
CICLing-2012, 13^{th} International Conference on Intelligent Text Processing and Computational Linguistics
IBERAMIA 2012, 13th edition of the Ibero-American Conference on Artificial Intelligence
CERI 2012, *II Congreso Nacional de Recuperación de la Información*
PAN CLEF 2012, Uncovering Plagiarism, Authorship, and Social Software Misuse
IEKA 2011, Workshop on Information Extraction and Knowledge Acquisition
PAN CLEF 2011, 5th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse
STIL 2011 : 8th Brazilian Symposium in Information and Human Language Technology
RANLP-2011 Student Research Workshop
ACL-HLT 2011, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Area chair
Student Session of the ACL-HLT 2011, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Faculty Advisor
CICLing-2011, 12^{th} International Conference on Intelligent Text Processing and Computational Linguistics
CICLing-2010, 11^{th} International Conference on Intelligent Text Processing and Computational Linguistics
MICAI-2010, 9th Mexican International Conference on Artificial Intelligence
MICAI-2009, 8th Mexican International Conference on Artificial Intelligence
RANLP-2009, Recent Advances in Natural Language Processing
CICLing-2009, 10^{th} International Conference on Intelligent Text Processing and Computational Linguistics
MICAI-2008, 7th Mexican International Conference on Artificial Intelligence
CICLing-2008, 9^{th} International Conference on Intelligent Text Processing and Computational Linguistics

NAACL-HLT 2007, The Annual Conference of the North American Chapter of the Association for Computational Linguistics

Journal Reviewer

Artificial Intelligence Journal, 2020
Computational Linguistics, 2015, 2018
Computer Speech and Language, 2017
PLOS ONE Journal, 2017
Language Resources and Evaluation, LREV, 2017
Transaction of the Association of Computational Linguistics (TACL), 2014, 2016, 2017
Computación y Sistemas, 2017
Editorial Board Member for the Journal Artificial Intelligence Research (JAIR), 2016
Procesamiento de Lenguaje Natural Journal, Journal for the Spanish Society of Natural Language Processing, 2016
Language & Linguistics Compass, 2015
ACM Transactions on Asian Language Information Processing, 2015, 2016
IEEE Transactions on Knowledge and Data Engineering, 2014
IEEE Transactions on Audio, Speech and Language Processing, 2013, 2014, 2015, 2016
Artificial Intelligence in Medicine, 2013, 2014
Natural Language Engineering, 2012, 2013
ACM Transactions on Intelligent Systems and Technology, TIST, 2011
Language Resources and Evaluation, 2011
Transactions on Intelligent Systems and Technology, 2011
Concurrency and Computation: Practice and Experience, 2010
INFORMS Journal on Computing, 2010
Computación y Sistemas, Special issue: "Innovative applications of AI", 2008

Services to Government Agencies

Panel reviewer for the National Science Foundation 2021, 2020, 2019, 2017, 2016, 2015, 2014, 2013, 2010, 2009.
Ad-hoc reviewer for the National Science Foundation, 2022, 2008.

Other Service Activities

Project reviewer for the Austrian Science Fund (FWF), 2015

12. Publications

Google Scholar Profile

Refereed Journal Articles

[J1] Siva Uday Sampreeth Chebolu, Paolo Rosso, Sudipta Kar, and **Thamar Solorio**. Survey on aspect category detection. *ACM Comput. Surv.*, may 2022. Just Accepted.

[J2] A. Pastor López-Monroy, Fabio A.González, and **Thamar Solorio**. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, 140:112909, 2020.

[J3] John Arevalo, **Thamar Solorio**, Manuel Montes y Gomez, and Fabio Gonzalez. Gated multimodal networks. *Neural Computing and Applications*, 2020.

[J4] John D. Osborne, Matthew B. Neu, Maria I. Danila, **Thamar Solorio**, and Steven J. Bethard. Cuiless2016: a clinical corpus applying compositional normalization of text mentions. *Journal of Biomedical Semantics*, 9(2), January 2018.

[J5] Dasha Bogdanova, Paolo Rosso, and **Thamar Solorio**. Exploring high-level features for detecting cyberpedophilia. *Comput. Speech Lang.*, 28(1):108–120, January 2014.

[J6] **Thamar Solorio**. Survey on emerging research on the use of natural language processing in clinical language assessment of children. *Language and Linguistics Compass*, 7(12):633–646, 2013.

[J7] Kairuh nisa Hassanali, Yang Liu, Aquiles Iglesias, **Thamar Solorio**, and Chrsitine Dollaghan. Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods*, 45(13), June 2013.

[J8] Gabriela Ramírez de-la Rosa, Manuel Montes y Gómez, **Thamar Solorio**, and Luis Villaseñor-Pineda. A document is known by the company it keeps: Neighborhood consensus for short text categorization. *Language Resources and Evaluation*, (47):127–149, 2012.

[J9] Keyur Gabani, **Thamar Solorio**, Yang Liu, and Christine Dollaghan. Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children. *Artificial Intelligence in Medicine*, 53(9):161–170, November 2011.

[J10] **Thamar Solorio**, Melissa Sherman, Yang Liu, Lisa Bedore, Elizabeth Peña, and Aquiles Iglesias. Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17(11):367–395, 2011.

[J11] M. Taufer, M-Y. Leung, **Thamar Solorio**, A. Licon, D. Mireles, D. Gomez-Leon, R. Araiza, and K.K. Johnson. RNAVLab: A unified environment for computational RNA structure analysis based on grid computing technology. *Parallel Computing*, 34:661–680, 2008.

[J12] Nigel G. Ward, Rafael Escalante, Yaffa Al Bayyari, and **Thamar Solorio**. Learning to show you're listening. *Computer Assisted Language Learning*, 20(4):385–407, 2007.

[J13] **Thamar Solorio**, Olac Fuentes, Roberto Terlevich, and Elena Terlevich. An active instance-based machine learning method for stellar population studies. *Monthly Notices of the Royal Astronomical Society*, 363(2):543–554, October 2005.

Peer Reviewed Conference and Workshop Papers

[C1] Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and **Thamar Solorio**. From none to severe: Predicting severity in movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3951–3956, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[C2] Farah Naz Chowdhury, Raga Shalini Koka, Mohammad Rajiur Rahman, **Thamar Solorio**, and Jaspal Subhlok. Identifying keyword predictors in lecture video screen text. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 281–286, 2021.

[C3] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, **Thamar Solorio**, Munmun De Choudhury, and Pamela J. Wisniewski. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[C4] Anjani Dhrangadhariya, Gustavo Aguilar, **Thamar Solorio**, Roger Hilfiker, and Henning Müller. End-to-end fine-grained neural entity recognition of patients, interventions, outcomes. In K. Selçuk Candan, Bogdan Ionescu, Lorraine Goeuriot, Birger Larsen, Henning Müller, Alexis Joly, Maria Maistro, Florina Piroi, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 65–77, Cham, 2021. Springer International Publishing.

[C5] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and **Thamar Solorio**. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[C6] Mahsa Shafaei, Christos Smailis, Ioannis Kakadiaris, and **Thamar Solorio**. A case study of deep learning-based multi-modal methods for labeling the presence of questionable content in movie trailers. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1297–1307, Held Online, September 2021. INCOMA Ltd.

[C7] Amirreza Shirani, Giai Tran, Hieu Trinh, Franck Dernoncourt, Nedim Lipka, Jose Echevarria, **Thamar Solorio**, and Paul Asente. PSED: A dataset for selecting emphasis in presentation slides. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4314–4320, 2021.

[C8] Dwija Parikh and **Thamar Solorio**. Normalization and back-transliteration for code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124, Online, June 2021. Association for Computational Linguistics.

[C9] Amirreza Shirani, Giai Tran, Hieu Trinh, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and **Thamar Solorio**. Learning to emphasize: Dataset and shared task models for selecting emphasis in presentation slides. *arXiv preprint arXiv:2101.03237*, 2021.

[C10] Shuguang Chen, Leonardo Neves, and **Thamar Solorio**. Mitigating temporal-drift: A simple approach to keep NER models crisp. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 163–169, Online, June 2021. Association for Computational Linguistics.

[C11] Sudipta Kar, Gustavo Aguilar, Mirella Lapata, and **Thamar Solorio**. Multi-view story characterization from movie plot synopses and reviews. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5629–5646, Online, November 2020. Association for Computational Linguistics.

[C12] Amirreza Shirani, Franck Dernoncourt, Jose Echevarria, Paul Asente, Nedim Lipka, and **Thamar Solorio**. Let me choose: From verbal context to font selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8607–8613, Online, July 2020. Association for Computational Linguistics. [PDF].

[C13] Gustavo Aguilar and **Thamar Solorio**. From English to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044, Online, July 2020. Association for Computational Linguistics. [PDF].

[C14] Gustavo Aguilar, Sudipta Kar, and **Thamar Solorio**. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France, May 2020. European Language Resources Association. [PDF].

[C15] Mahsa Shafaei, Niloofar Safi Samghabadi, Sudipta Kar, and **Thamar Solorio**. Age suitability rating: Predicting the MPAA rating based on movie dialogues. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1327–1335, Marseille, France, May 2020. European Language Resources Association. [PDF].

[C16] Mahsa Shafaei, Adrian Pastor Lopez-Monroy[p], and **Thamar Solorio**. Exploiting textual, visual and product features for predicting the likeability of movies. In *32nd Florida Artificial Intelligence Conference (FLAIRS-32)*. The 32nd International FLAIRS Conference, 2019. [PDF].

[C17] Amirreza Shirani, Bowen Xu, David Lo, **Thamar Solorio**, and Amin Alipour. Question relatedness on stack overflow: The task, dataset, and corpus-inspired models. In *AAAI Reasoning for Complex Question Answering Workshop (AAAI 2019)*, 2019. [PDF].

[C18] Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and **Thamar Solorio**. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, Florence, Italy, July 2019. Association for Computational Linguistics. [PDF].

[C19] Suraj Maharjan, Deepthi Mave, Prasha Shrestha, Manuel Montes, Fabio A. González, and **Thamar Solorio**. Jointly learning author and annotated character n-gram embeddings: A case study in literary text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 684–692, Varna, Bulgaria, September 2019. INCOMA Ltd. [PDF].

[C20] Amirreza Shirani, Pastor Lopez-Monroy^p, Fabio Gonzalez, **Thamar Solorio**, and Mohammad Amin Alipour. Evaluation of type inference with textual cues. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018. [PDF].

[C21] Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and **Thamar Solorio**. Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia, July 2018. Association for Computational Linguistics. [PDF].

[C22] Sudipta Kar, Suraj Maharjan, and **Thamar Solorio**. Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 2879–2891, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. [PDF].

[C23] Deepthi Mave, Suraj Maharjan, and **Thamar Solorio**. Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia, July 2018. Association for Computational Linguistics. [PDF].

[C24] Suraj Maharjan, Manuel Montes, Fabio A. González, and **Thamar Solorio**. A genre-aware attention model to improve the likability prediction of books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3381–3391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. [PDF].

[C25] Adrian Pastor López-Monroy^p, Fabio A. González, Manuel Montes, Hugo Jair Escalante, and **Thamar Solorio**. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long Papers)*, pages 1216–1225, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. [PDF].

[C26] Gustavo Aguilar, Adrian Pastor López-Monroy^p, Fabio González, and **Thamar Solorio**. Modeling noisiness to recognize named entities using multitask neural networks on social media. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long Papers)*, pages 1401–1412, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. [PDF].

[C27] Suraj Maharjan, Sudipta Kar, Manuel Montes, Fabio A. González, and **Thamar Solorio**. Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 259–265, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. [PDF].

[C28] Sudipta Kar, Suraj Maharjan, A. Pastor López-Monroy^p, and **Thamar Solorio**. MPST: A corpus of movie plot synopses with tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). [PDF].

[C29] Thoudam Doren Singh^p and **Thamar Solorio**. Towards translating mixed-code comments from social media. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 457–468, Cham, 2018. Springer International Publishing.

[C30] John Arevalo, **Thamar Solorio**, Manuel Montes y Gómez, and Fabio Gonzalez. Gated multimodal units for information fusion. In *Workshop Track -International Conference on Learning Representations (ICLR-2017)*, Toulon, France, April 2017. [PDF].

[C31] Prasha Shrestha, Sebastian Sierra, Fabio A. Gonzalez, Manuel Montes y Gómez, and **Thamar Solorio**. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the EACL*, Valencia, Spain, 2017. EACL. [PDF].

[C32] Suraj Maharjan, John Arevalo, Fabio A. Gonzalez, Manuel Montes y Gómez, and **Thamar Solorio**. A multi-task approach to predict likability of books. In *Proceedings of the EACL*, Valencia, Spain, 2017. EACL. [PDF].

[C33] Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and **Thamar Solorio**. Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics.

[C34] Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and **Thamar Solorio**. Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas, November 2016. Association for Computational Linguistics. [PDF].

[C35] Fahad AlGhamdi, Giovanni Molina, Mona Diab, **Thamar Solorio**, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Austin, Texas, November 2016. Association for Computational Linguistics. [PDF].

[C36] Upendra Sapkota, **Thamar Solorio**, Manuel Montes y Gómez, and Steven Bethard. Domain adaptation for authorship attribution: Improved structural correspondence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2226–2235, Berlin, Germany, August 2016. ACL, Association for Computational Linguistics. [PDF].

[C37] Prasha Shrestha, Arjun Mukherjee, and **Thamar Solorio**. Large scale authorship attribution of online reviews. In *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016*, Konya, Turkey, 2016.

[C38] Prasha Shrestha, Nicolas Rey-Villamizar, Farig Sadeque, Ted Pedersen, Steven Bethard, and **Thamar Solorio**. Age and gender prediction on health forum data. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *10th Edition of the Language Resources and Evaluation Conference, LREC*. European Language Resources Association (ELRA), 2016. [PDF].

[C39] Nicolas Rey-Villamizar, Prasha Shrestha, Farig Sadeque, Steven Bethard, Ted Pedersen, Arjun Mukherjee, and **Thamar Solorio**. Analysis of anxious word usage on online health forums. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 37–42, Auxtin, TX, November 2016. Association for Computational Linguistics. [PDF].

[C40] Farig Sadeque, Ted Pedersen, **Thamar Solorio**, Prasha Shrestha, Nicolas Rey-Villamizar, and Steven Bethard. Why do they leave: Modeling participation in online depression forums. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 14–19, Austin, TX, USA, November 2016. Association for Computational Linguistics. [PDF].

[C41] Farig Sadeque, **Thamar Solorio**, Ted Pedersen, Prasha Shrestha, and Steven Bethard. Predicting continued participation in online health forums. In *Proceedings of the 6th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 15–20, Lisboa, Portugal, 2015. Association for Computational Linguistics. [PDF].

[C42] R. Verma, M. Kantarcioglu, D. Marchette, E. Leiss, and **Thamar Solorio**. Security analytics: Essential data analytics knowledge for cybersecurity professionals and students. *IEEE Security Privacy*, 13(6):60–65, 2015.

[C43] Suraj Maharjan, Elizabeth Blair, Steven Bethard, and **Thamar Solorio**. Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA, 2015. ACL. [PDF].

[C44] Upendra Sapkota, Steven Bethard, Manuel Montes y Gómez, and **Thamar Solorio**. Not all character n-grams are created equal: A study in authorship attribution. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, pages 93–102, Denver, Colorado, 2015. ACL. [PDF].

[C45] Prasha Shrestha and **Thamar Solorio**. Identification of original document by using textual similarities. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, pages 643–654, 2015.

[C46] Suraj Maharjan, Prasha Shrestha, **Thamar Solorio**, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In *Proceedings of the 14th Iberoamerican Conference on AI*, pages 62–72, Santiago de Chile, Chile, November 2014. Advances in Artificial Intelligence- Lecture Notes in Computer Science. [PDF]$^{/UH}$.

[C47] Prasha Shrestha, Suraj Maharjan, **Thamar Solorio**, and Ragib Hasan. Using string information for malware family identification. In *Proceedings of the 14th Iberoamerican Conference on AI*, pages 686–697, Santiago de Chile, Chile, November 2014. Advances in Artificial Intelligence- Lecture Notes in Computer Science. [PDF]$^{/UH}$.

[C48] **Thamar Solorio**, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October 2014. Association for Computational Linguistics. [PDF]$^{/UH}$.

[C49] Upendra Sapkota, **Thamar Solorio**, Manuel Montes, Steven Bethard, and Paolo Rosso. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. [PDF].

[C50] **Thamar Solorio**, Ragib Hasan, and Mainul Mizan. Sockpuppet detection in wikipedia: A corpus of real-world deceptive writing for linking identities. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 26–31, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). arXiv preprint arXiv:1310.6772,[PDF].

[C51] Gabriela Ramírez de-la Rosa, **Thamar Solorio**, Manuel Montes-y-Gómez, Yang Liu, Aquiles Iglesias, Lisa Bedore, and Elizabeth Peña. Exploring word class n-grams to measure language development in children. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 89–97, Sofia, Bulgaria, August 2013. ACL. [PDF].

[C52] Khairun nisa Hassanali, Yang Liu, and **Thamar Solorio**. Using latent dirichlet allocation for child narrative analysis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 111–115, Sofia, Bulgaria, August 2013. ACL. [PDF].

[C53] **Thamar Solorio**, Ragib Hasan, and Mainul Mizan. A case study of sockpuppet detection in wikipedia. In *Workshop on Language Analysis in Social Media (LASM) at NAACL-HLT 2013*, pages 59–68, Atlanta, Georgia, June 2013. ACL. [PDF].

[C54] Upendra Sapkota, **Thamar Solorio**, Manuel Montes-y-Gómez, and Paolo Rosso. The use of orthogonal similarity relations in the prediction of authorship. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2013*, pages 463–475, Samos, Greece, March 2013.

[C55] Khairun nisa Hassanali, Yang Liu, and **Thamar Solorio**. Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. In *Proceedings of 3rd Workshop on Child, Computer and Interaction (WOCCI 2012)*, 2012. [PDF].

[C56] Khairun nisa Hassanali, Yang Liu, and **Thamar Solorio**. Evaluating NLP features for automatic prediction of language impairment using child speech transcripts. In *Proceedings of INTERSPEECH 2012*, 2012.

[C57] Daria Bogdanova, Paolo Rosso, and **Thamar Solorio**. On the impact of sentiment and emotion based features in detecting online sexual predators. In *Proceedings of the ACL 2012 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 110–118, Jeju, Republic of Korea, July 2012. ACL. [PDF].

[C58] Binod Gyawali, **Thamar Solorio**, and Yassine Benajiba. Grading the quality of medical evidence. In *2012 Workshop on Biomedical Natural Language Processing (BIONLP 2012)*, pages 176–184, Montréal, Canada, June 2012. ACL. [PDF].

[C59] Daria Bogdanova, Paolo Rosso, and **Thamar Solorio**. Modelling fixated discourse in chats with cyberpedophiles. In *EACL 2012 Workshop on Computational Approaches to Deception Detection*, pages 86–90, Avignon, France, April 2012. ACL.

[C60] Gabriela Ramírez de-la Rosa, **Thamar Solorio**, Manuel Montes-y-Gómez, Yang Liu, Lisa Bedore, Elizabeth Peña, and Aquiles Iglesias. Language dominance prediction in Spanish-English bilingual children using syntactic information: a first approximation. In *ICL Workshop on Iberian Cross-Language NLP tasks*, Huelva, Spain, September 2011.

[C61] Hugo Jair Escalante, Manuel Montes-y-Gómez, and **Thamar Solorio**. Weighted profile intersection measure for profile-based authorship attribution. In *10th Mexican International Conference on Artificial Intelligence*, pages 232–243, Puebla, Mexico, November 2011.

[C62] Debangana Dey, **Thamar Solorio**, Manuel Montes-y-Gómez, and Hugo Escalante. Instance selection based on the silhouette coefficient measure for text classification. In *10th Mexican International Conference on Artificial Intelligence*, pages 357–369, Puebla, Mexico, November 2011.

[C63] Binod Gyawali, **Thamar Solorio**, Manuel Montes-y-Gómez, Brad Wardman, and Gary Warner. Evaluating a semisupervised approach to phishing url identification in a realistic scenario. In *8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS 2011)*, pages 176–183, Perth, Australia, 2011. ACM. [PDF].

[C64] **Thamar Solorio**, Sangita Pillay, Sindhu Raghavan, and Manuel Montes-y-Gómez. Modality specific metafeatures for authorship attribution on web forum posts. In *5th International Joint Conference on Natural Language Processing, IJCNLP 2011*, pages 156–164, Chiang Mai, Thailand, November 2011. AFNLP. [PDF].

[C65] H. Jair Escalante, **Thamar Solorio**, and Manuel Montes-y-Gómez. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting for the Association for Computational Linguistics: Human Language Technologies*, pages 288–298, Portland, Oregon, June 19–24 2011. ACL. [PDF].

[C66] Aaron Blum, Brad Wardman, **Thamar Solorio**, and Gary Warner. Lexical feature based phishing url detection using online learning. In *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security*, AISec '10, pages 54–60, Chicago, Illinois, 2010. ACM. [PDF].

[C67] Sangita Pillay and **Thamar Solorio**. Authorship attribution on web forum posts. In *eCrime Researchers Summit*, Dallas, TX, October 2010. APWG. [PDF].

[C68] Gabriela Ramírez de-la Rosa, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, David Pinto-Avendaño, and **Thamar Solorio**. Using information from the target language to improve crosslingual text classification. In *Proceedings of IceTAL 2010*, pages 305–313, Reykjavik, Iceland, August 2010.

[C69] Richa Tiwari, Chengcui Zhang, and **Thamar Solorio**. A supervised machine learning approach of extracting coexpression relationship among genes from literature. In *Proceedings of IEEE IRI 2010*, pages 98–103, Las Vegas, NV, 2010.

[C70] Keyur Gabani, Melissa Sherman, **Thamar Solorio**, Yang Liu, Lisa Bedore, and Elizabeth Peña. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT) 2009*, pages 46–55, Boulder, Colorado, June 2009. ACL. [PDF].

[C71] **Thamar Solorio** and Yang Liu. Part-of-speech tagging for English-Spanish code-switched text. In *Empirical Methods on Natural Language Processing, EMNLP-2008*, pages 1051–1060, Honolulu, Hawaii, October 2008. ACL. [PDF].

[C72] **Thamar Solorio** and Yang Liu. Learning to predict code-switching points. In *Empirical Methods on Natural Language Processing, EMNLP-2008*, pages 973–981, Honolulu, Hawaii, October 2008. ACL. [PDF].

[C73] **Thamar Solorio** and Yang Liu. Using language models to identify language impairment in Spanish-English bilingual children. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 116–117, Columbus, Ohio, June 2008. ACL. [PDF].

[C74] Michela Taufer, **Thamar Solorio**, Abel Licon, David Mireles, and Ming-Ying Leung. On the effectiveness of rebuilding RNA secondary structures from sequence chunks. In *The Seventh IEEE International Workshop on High-Performance Computational Biology, HiCOMB 2008*, Miami, Florida, April 2008.

[C75] Olac Fuentes, David Vera, and **Thamar Solorio**. A filter-based approach to detect end-of-utterances from prosody in dialog systems. In *Proceedings of the The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 45–48, Rochester, NY, April 2007. ACL. [PDF].

[C76] Juan C. Franco and **Thamar Solorio**. Baby steps towards a language model for Spanglish. In Alexander Gelbukh, editor, *The Eighth International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2007*, volume 4394 of *Lecture Notes in Computer Science*, pages 75–84, Mexico City, February 2007.

[C77] **Thamar Solorio**, Olac Fuentes, Nigel Ward, and Yaffa Al Bayyari. Prosodic feature generation for back-channel prediction. In *The Ninth International Conference on Spoken Language Processing, INTERSPEECH 2006*, Pittsburgh, Pennsylvania, September 2006.

[C78] Ted Pedersen, Anagha Kulkarni, Roxana Angheluta, Zornista Kozareva, and **Thamar Solorio**. Improving name discrimination : A language salad approach. In *Proceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction*, pages 25–32, April 2006. [PDF].

[C79] Ted Pedersen, Anagha Kulkarni, Roxana Angheluta, Zornista Kozareva, and **Thamar Solorio**. An unsupervised language independent method of name discrimination using second order co-occurrence features. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Seventh International Conference, CICLing 2006*, volume 3878 of *Lecture Notes in Computer Science*, pages 208–222, February 2006.

[C80] **Thamar Solorio**. Exploiting named entity taggers in a second language. In Chris Callison-Burch and Stephen Wan, editors, *Student Research Workshop at the* 43^{rd} *Annual Meeting of the Association for Computational Linguistics, ACL-2005*, pages 25–30, Ann Arbor, Michigan, June 2005. [PDF].

[C81] **Thamar Solorio** and Aurelio López López. Learning named entity recognition in Portuguese from Spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Sixth International Conference, CICLing 2005*, volume 3406 of *Lecture Notes in Computer Science*, pages 762–768, February 2005.

[C82] **Thamar Solorio**, Manuel Pérez Coutiño, Manuel Montes y Gómez, Luis Villaseñor Pineda, and Aurelio López López. Question classification in Spanish and Portuguese. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Sixth International Conference, CICLing 2005*, volume 3406 of *Lecture Notes in Computer Science*, pages 612–619, February 2005.

[C83] Olac Fuentes, **Thamar Solorio**, Roberto Terlevich, and Elena Terlevich. Analysis of galactic spectra using active instance-based learning and domain knowledge. In Christian Lemaître, Carlos Reyes, and Jesús A. González, editors, *Advances in Artificial Intelligence – IBERAMIA 2004: The IX Ibero-American Conference on Artificial Intelligence*, volume 3315 of *Lecture Notes in Artificial Intelligence 3315*, pages 215–224, Puebla, Mexico, November 2004.

[C84] Manuel Pérez Coutiño, **Thamar Solorio**, Manuel Montes y Gómez, Aurelio López López, and Luis Villaseñor Pineda. Question answering for Spanish based on lexical and context annotation. In Christian Lemaître, Carlos Reyes, and Jesús A. González, editors, *Advances in Artificial Intelligence – IBERAMIA 2004*, volume 3315 of *Lecture Notes in Artificial Intelligence 3315*, pages 325–333, Puebla, Mexico, November 2004.

[C85] Manuel Pérez-Coutiño, **Thamar Solorio**, Manuel Montes y Gómez, Aurelio López López, and Luis Villaseñor Pineda. The use of lexical context in question answering for Spanish. In Carol Peters and Francesca Borri, editors, *Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004)*, Bath, England, September 2004.

[C86] **Thamar Solorio**, Manuel Pérez Coutiño, Manuel Montes y Gómez, Luis Villaseñor Pineda, and Aurelio López López. A language independent method for question classification. In *The 20th International*

Conference on Computational Linguistics, COLING-04, volume II, pages 1374–1380, Geneva, Switzerland, August 2004. [PDF].

[C87] Manuel Pérez Coutiño, **Thamar Solorio**, Manuel Montes y Gómez, Aurelio López López, and Luis Villaseñor Pineda. Toward a document model for question answering systems. In Jesus Favela, Ernestina Menasalvas, and Edgar Chávez, editors, *Advances in Web Intelligence: Second International Atlantic Web Intelligent Conference AWIC 2004*, volume 3034 of *Lecture Notes in Artificial Intelligence 3034*, pages 145–154, Cancun, Mexico, May 2004.

[C88] Olac Fuentes and **Thamar Solorio**. An optimization algorithm based on active and instance-based learning. In R. Monroy, G. Arroyo-Figueroa, L. E. Sucar, and H. Sossa, editors, *MICAI 2004: Advances in Artificial Intelligence, Third Mexican International Conference on Artificial Intelligence*, volume 2972 of *Lecture Notes in Artificial Intelligence 2972*, pages 242–251, Mexico City, Mexico, April 2004.

[C89] **Thamar Solorio** and Aurelio López López. Learning named entity classifiers using support vector machines. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Fifth International Conference, CICLing 2004*, volume 2945/2004 of *Lecture Notes in Computer Science 2945*, pages 158–167, February 2004.

[C90] **Thamar Solorio** and Olac Fuentes. Improving classification accuracy of large test sets using the ordered classification algorithm. In F.J. Garijo, J.C. Riquelme, and M. Toro (Eds.), editors, *Advances in Artificial Intelligence - IBERAMIA 2002: 8th Ibero-American Conference on Artificial Intelligence*, volume 2527 of *Lecture Notes in Computer Science 2527*, pages 70–79, Seville, Spain, November 2002.

Preprint Papers

[P1] Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and **Thamar Solorio**. Survey of aspect-based sentiment analysis datasets, 2022.

[P2] Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona Diab, and **Thamar Solorio**. Calcs 2021 shared task: Machine translation for code-switched data, 2022.

[P3] Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and **Thamar Solorio**. Exploring conditional text generation for aspect-based sentiment analysis. 2021.

[P4] Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and **Thamar Solorio**. From none to severe: Predicting severity in movie scripts, 2021.

[P5] **Thamar Solorio**, Mahsa Shafaei, Christos Smailis, Brad J. Bushman, Douglas A. Gentile, Erica Scharrer, Laura Stockdale, and Ioannis Kakadiaris. White paper – objectionable online content: What is harmful, to whom, and why, 2021.

[P6] **Thamar Solorio**, Mahsa Shafaei, Christos Smailis, Mona Diab, Theodore Giannakopoulos, Heng Ji, Yang Liu, Rada Mihalcea, Smaranda Muresan, and Ioannis Kakadiaris. White paper: Challenges and considerations for the creation of a large labelled repository of online videos with questionable content, 2021.

[P7] Amirreza Shirani, Giai Tran, Hieu Trinh, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and **Thamar Solorio**. Learning to emphasize: Dataset and shared task models for selecting emphasis in presentation slides, 2021.

[P8] **Thamar Solorio**, Mahsa Shafaei, Christos Smailis, Isabelle Augenstein, Margaret Mitchell, Ingrid Stapf, and Ioannis Kakadiaris. White paper-creating a repository of objectionable online content: Addressing undesirable biases and ethical considerations.

Shared Task Papers

[T1] Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, **Thamar Solorio**, and Amitava Das. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[T2] Amirreza Shirani, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and **Thamar Solorio**. SemEval-2020 task 10: Emphasis selection for written text in visual media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1360–1370, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[T3] Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and **Thamar Solorio**. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France, May 2020. European Language Resources Association (ELRA). [PDF].

[T4] Niloofar Safi Samghabadi, Deepthi Mave, Sudipta Kar, and **Thamar Solorio**. RiTUAL-UH at TRAC 2018 shared task: Aggression identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 12–18, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. 🏆1st Place-Hindi Facebook Task, 2nd Place- Hindi Social Medial Task 🏆 [PDF].

[T5] Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and **Thamar Solorio**. Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia, July 2018. Association for Computational Linguistics. [PDF].

[T6] Gustavo Aguilar, Suraj Maharjan, Pastor Lopez-Monroy^p, Fabio A. Gonzalez, and **Thamar Solorio**. A multi-task approach for named entity recognition in social media data. In *Proceedings of EMNLP-2017 Workshop on Noisy User Generated data (W-NUT)*, Copenhagen, Denmark, 2017. 🏆1st Place-WNUT 2017 Shared Task on Novel and Emerging Named Entity Recognition 🏆[PDF].

[T7] Sebastian Sierra, Manuel Montes-Y-Gómez, **Thamar Solorio**, and Fabio A. González. Convolutional Neural Networks for Author Profiling in PAN 2017—Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*. CEUR-WS.org, September 2017. [PDF].

[T8] Adrián Pastor Lopez-Monroy, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Thamar Solorio. Social-Media Users can be Profiled by their Similarity with other Users—Notebook for PAN at CLEF 2017. In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*. CEUR-WS.org, September 2017. [PDF].

[T9] Sudipta Kar, Suraj Maharjan, and **Thamar Solorio**. RiTUAL-UH at SemEval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 877–882, Vancouver, Canada, August 2017. Association for Computational Linguistics. 🏆2nd place Sub-Task 2 SemEval-2017 Task-5 on Fine-Grained Sentiment Analysis on Financial Microblogs and News 🏆 [PDF].

[T10] Mohammed Attia, Suraj Maharjan, Younes Samih, Laura Kallmeyer, and **Thamar Solorio**. CogALex-V shared task: GHHH - detecting semantic relations via word embeddings. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 86–91, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. 🏆1st Place Task 1, 2nd Place Task 2, CogALex-2016 Shared Task on Corpus-Based Identification of Semantic Relations 🏆 [PDF].

[T11] Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and **Thamar Solorio**. Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas, November 2016. Association for Computational Linguistics. [PDF].

[T12] Nicolas Rey-Villamizar, Prasha Shrestha, **Thamar Solorio**, Farig Sadeque, Steven Bethard, and Ted Pedersen. Semi-supervised CLPsych 2016 shared task system submission. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 171–175, San Diego, CA, USA, June 2016. Association for Computational Linguistics. [PDF].

[T13] Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and **Thamar Solorio**. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas, November 2016. Association for Computational Linguistics. [PDF].

[T14] Marc Franco-Salvador, Sudipta Kar, **Thamar Solorio**, and Paolo Rosso. UH-PRHLT at SemEval-2016 task 3: Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 814–821, San Diego, California, June 2016. Association for Computational Linguistics. 🏆1st Place Task B, SemEval 2016 Task 3: Community Question Answering 🏆 [PDF].

[T15] Suraj Maharjan and **Thamar Solorio**. Using a wide range of fetures for author profiling. In *Notebook for PAN at CLEF 2015 Evaluation Labs and Workshop*, Toulouse, France, September 2015. [PDF].

[T16] Suraj Maharjan, Prasha Shrestha, and **Thamar Solorio**. A simple approach to author profiling in mapreduce. In *Notebook for PAN at CLEF 2014*, Sheffield, UK, 2014.

[T17] Prasha Shrestha, Suraj Maharjan, and **Thamar Solorio**. Machine translation evaluation metric for text alignment. In *Notebook for PAN at CLEF 2014*, Sheffield, UK, 2014.

[T18] John David Osborne, Binod Gyawali, and **Thamar Solorio**. Evaluation of freely available open source software for clinical concept recognition. In *Notebook for ShARe/CLEF eHealth Evaluation Lab at CLEF 2013*, 2013.

[T19] Binod Gyawali, Gabriela Ramirez de-la Rosa, and **Thamar Solorio**. Native language identification: a simple n-gram based approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–231, Atlanta, Georgia, June 2013. ACL.

[T20] Prasha Shrestha and **Thamar Solorio**. Using a variety of n-grams for the detection of different kinds of plagiarism. In *Notebook for PAN at CLEF 2013*, 2013.

[T21] Upendra Sapkota, **Thamar Solorio**, Manuel Montes-y-Gómez, and Gabriela Ramírez de-la Rosa. Automatic author profiling for English and Spanish text. In *Notebook for PAN at CLEF 2013*, 2013.

[T22] Binod Gyawali and **Thamar Solorio**. UABCoRAL: A preliminary study for resolving the scope of negation. In **SEM: 1st Joint Conference on Lexical and Computational Semantics*, pages 275–281, Montréal, Canada, June 2012. ACL.

[T23] Upendra Sapkota and **Thamar Solorio**. Sub-profiling by linguistic dimensions to solve the authorship attribution task. In *Notebook for PAN at CLEF 2012*, Rome, Italy, September 2012.

[T24] **Thamar Solorio**, Sangita Pillay, and Manuel Montes-y-Gómez. Authorship identification with modality specific meta features. In *Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2011) held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, Amsterdam, September 2011.

13. Teaching Experience

Advanced NLP: Deep Learning for Natural Language Understanding (graduate), UH Fall 2020, Fall 2017
Computer Science and Programming (undergraduate). UH Spring 2015, 2016, 2017, 2018, Fall 2020
Computational Thinking (undergraduate). UAB Spring 2014
Intro to Object Oriented Programming in Java (undergraduate). UAB Spring 2013, 2014, Fall 2013
Natural Language Processing (graduate). UAB Spring 2010, 2011, UH Fall 2015, Spring 2018, Spring 2020
Artificial Intelligence (graduate). UAB Fall 2009, 2010, Spring 2012, 2013
Fluency in Information Technology (undergraduate). UAB Fall 2010, 2011, Spring 2012
Elementary Data Structures and Algorithms (undergraduate). UTEP Fall 2005, 2006, Spring 2006, 2007
Computer Programming for Science & Engineering (undergraduate). UTEP Fall 2006, Spring 2007
Artificial Intelligence (undergraduate). UTEP Fall 2006

Software Engineering I (undergraduate). UTEP Fall 2005, Spring 2006
Computer Architecture I (undergraduate). UTEP Fall 2005
Selected Topics in Artificial Intelligence (graduate). Instituto Tecnológico de Apizaco, 2003
Natural Language Processing (graduate). Instituto Tecnológico de Apizaco, 2003
Object Oriented Programming in Java (undergraduate). Universidad Popular Autónoma del Estado de Puebla 2002
Teaching Assistant, Automata Theory and Formal Languages (graduate). Instituto Nacional de Astrofísica, Óptica y Electrónica, assistant for professor Aurelio López, 2004
Teaching Assistant, Machine Learning (graduate) . Instituto Nacional de Astrofísica, Óptica y Electrónica, assistant for professor Olac Fuentes, 2004
Teaching Assistant, Research Seminar II (graduate). Instituto Nacional de Astrofísica, Óptica y Electrónica, assistant for professor Olac Fuentes, 2003
Teaching Assistant, Automata Theory and Formal Languages (graduate). Instituto Nacional de Astrofísica, Óptica y Electrónica, graduate level, assistant for professor Ariel Carrasco, 2001

14. Professional Affiliations

- Association for Computing Machinery (ACM)
- Association for Computational Linguistics (ACL)

15. Postdoctoral Mentoring

Pastor López-Monroy (2017-2018)
Thoudam Doren Singh (2015-2016)

16. Student Advising and Mentoring

Current Ph.D. Students
Shuguang Chen
Siva Sampreeth Chebolu
Yigeng Zhang

Advised PhD Graduates
Gustavo Aguilar Alas, Applied Scientist, Alexa AI, (2020 UHCS Best PhD Student, 2019 Snap Fellow, 2018 UHCS Best Junior PhD Award)
John Arevalo, Postdoctoral Researchers, Broad Institute of MIT and Harvard (co-advised with Prof. Fabio Gonzalez)
Sudipta Kar, Applied Scientist, Alexa AI (2019 UHCS Best PhD Student)
Suraj Maharjan, Applied Scientist, Alexa AI
Niloofar Safi Samghabadi, Research Scientist, Expedia Group, CRA-W 2016 and GHC 2017 travel grant recipient
Upendra Sapkota, Data Scientist, Apple (co-advised with Prof. Steven Bethard)
Mahsa Shafaei, Microsoft Research, 2018 Grace Hopper Celebration of Women in Computing travel grant recipient
Amirreza Shirani, Apple
Prasha Shrestha, Applied Scientist, Amazon

Advised MS Graduates
Debangana Dey, M.S., UAB 2012 (2011 ACM-W travel scholarship recipient)
Maksim Egorov, UH 2019
Binod Gyawali, M.S. UAB 2013, now Research Engineer at ETS
Akshay Kulkarni, UH 2017-2018
Deepthi Mave, UH 2019
María Petra Paredes, Sumarización Automática de Documentos, M.S in Computer Science, Instituto Tecnológico de Apizaco, 2004.

Sangita Pillay, M.S., UAB 2011 (2010 ACM-W travel scholarship recipient)
Gabriela Ramirez de la Rosa, M.S. UAB 2013, 2013 CRA-W travel award recipient, now Faculty at the Universidad Autónoma Metropolitana, Cuajimalpa, Mexico
Melissa Sherman, M.S., UTD 2009

Current Undergraduate Student Advising

Giai Nguyen, UH
Dwija Parikh, 2020 UH SURF Scholarship recipient

Advised BS Graduates
Kenny G. Blackmon, UAB 2011
Elizabeth Blair, UAB 2013-2014
Cash deLeon, UH
Anthony Bowman, UAB 2010
Salim El Awad, UH 2015
Keith Erdbruegger, UH 2016-2017
Juan Carlos Franco, UTEP
Afsheen Hatami, UH, 2019 NLM Summer Undergraduate Program
Wesley Johnson, UAB 2010
Brenda Medina, UTEP, (CRA-W DMP Student Awardee 2008)
David Mireles, UTEP, (Microsoft Scholar, Google Scholar, DSSI-2007 Fellow)
Armando Morales, UH
Kevin Portillo, UH 2016
Raymund Riegl, UH 2015-2016
Aidaly Santamaria, UH (2015)
Cameron Stanley, UAB 2011
Simon Tice, UH 2016
David Vera, UTEP
Brian Whitley, UAB 2012
Nicolle Whitman, UTEP, (LSMAP Scholar), now at Microsoft

PhD Thesis Committee Member

1. Fatih Akdag, UH
2. Miloud Aqqa, UH
3. Afiya Ayman, UH
4. Marzieh Berenjkoub, UH, 2019
5. Avisha Das, UH
6. Zekai Demirezen, UAB 2012
7. Ana Valeria Gonzalez, University of Copenhagen, 2021
8. Khairun-nisa Hassanali, UTD 2013
9. Myriam Hernandez, Universidad de Alicante, Spain, 2015
10. Daniel Lee, UH
11. Zhenggang Li, UH
12. Fabian Ng (Physics), UH
13. David O'Gwynn, UAB 2011

14. John Osborne, UAB
15. Rajiur Rahman, UH
16. Md Rezaul Karim Raju, UH
17. Carlos Romero, UH (Spanish Linguistics), 2016
18. Christos Smailis, UH
19. Salah Uddin, UH
20. Genta Indra Winata, The Hong Kong University of Science and Technology, 2021

MS Thesis Committee Member

1. Moahammad Rajiur Rahman, UH
2. Roga Shalini Koka, UH, 2019
3. Saba Khan, UH, 2020

16. News and Media Coverage

KHOU December 2018
UH student develops website to simplify picking a movie
URL: https://www.khou.com/article/news/uh-student-develops-website-to-simplify-picking-a-movie/285-624911544

New Scientist November 2013
Unmask Wikipedia sock puppets by the way they write
URL: https://www.newscientist.com/article/mg22029434.500-unmask-wikipedia-sock-puppets-by-the-way-they-write/

The Signpost, Wikipedia Newspaper June 2013
Sockpuppet evidence from automated writing style analysis
URL: https://en.wikipedia.org/wiki/Wikipedia:Wikipedia-Signpost/2013-06-26/Recent-research