Annual Report for Period: 08/2009 - 07/2010
Submitted on: 08/12/2010
Principal Investigator: Carbonell, Jaime G.
Award ID: 0534217
Organization: Carnegie Mellon University
Submitted By:
Lavie, Alon - Co-Principal Investigator
Title:
LETRAS: A Learning-based Framework for Machine Translation of Low Resource Languages

Project Participants

Senior Personnel

Name: Carbonell, Jaime
Worked for more than 160 Hours: Yes
Contribution to Project:
Jaime Carbonell is the PI of the project and oversees strategic planning and overall goals for the project. He co-advises two PhD students working on the project.

Name: Lavie, Alon
Worked for more than 160 Hours: Yes
Contribution to Project:
Alon Lavie is co-PI and is the main director of research in the project. He spearheads the research on the development of the Stat-XFER MT engine and its application to the development of MT prototype systems for resource-poor and resource-rich scenarios.
He advises or co-advises all students working on the project.

Name: Levin, Lorraine
Worked for more than 160 Hours: Yes
Contribution to Project:
Lori is an integral faculty member of our research group. Within LETRAS, Lori has primary oversight responsibilities for our collaboration with the University of Alaska - Fairbanks, where we are applying our MT approach to low-resource native Alaskan languages (initially Inupiaq). Lori spearheads our group's work
on data elicitation for MT, highly relevant to our work on LETRAS,
but currently funded by other sources. Lori is a PhD thesis committee member for Ariadna Font Llitjos - a PhD student funded by LETRAS.

Name: Frederking, Robert
Worked for more than 160 Hours: No
Contribution to Project:
Bob Frederking is an LTI faculty member affiliated with our research group. Bob co-advises a student working on feature detection and navigation (a closely related but separately funded project).

Post-doc

Graduate Student

Name: Font Llitjos, Ariadna
Worked for more than 160 Hours: Yes
Contribution to Project:
Ariadna Font Llitjos was the senior PhD student supported by the LETRAS grant. Her PhD research was on automatic rule refinement,
applied within our transfer-based MT approach. Ariadna successfully defended her thesis in mid July 2007, and left the project at the end of August 2007.

Name: Ambati, Vamshi

Worked for more than 160 Hours: Yes

Contribution to Project:

Vamshi is a new PhD student that joined our group in June 2007.
His research work is focusing on methods for automatically extracting transfer rules from Parallel data

Name: Agarwal, Abhaya

Worked for more than 160 Hours: Yes

Contribution to Project:

Abhaya Agarwal is a PhD student in our research group. Abhaya was the main student in charge of building our Urdu-to-English Stat-XFER MT system, which was entered into the 2008 NIST MT Evaluation.

Name: Hanneman, Gregory

Worked for more than 160 Hours: Yes

Contribution to Project:

Greg Hanneman is a PhD student in our research group. Greg participated in the research leading to our Urdu-to-English Stat-XFER MT system that was entered into the 2008 NIST MT Evaluation.
Greg was in charge of running our multi-engine system combination approach to combine Stat-XFER MT with Phrase-based SMT systems for Urdu.

Name: Clark, Jonathan

Worked for more than 160 Hours: Yes

Contribution to Project:

Jonathan Clark joined our research group in Fall 2007, but became integrally involved in research related to the MT developement under LETRAS in Fall 2008. He has taken the role of primary system maintainer, had a major role in development of our Czech-to-English Stat-XFER prototype for the MT Marathon in Prague in January 2009, supported the development of Stat-XFER systems for Hebrew and Portuguese, and has done significant work on migrating the Stat-XFER framework to teh new Joshua decoder.

Name: Parlikar, Alok

Worked for more than 160 Hours: No

Contribution to Project:

Alok Parlikar is a graduate student that is a member of our group. His primary role in the project was the development of several versions of our Urdu-to-English MT system that were submitted to the NIST-2009 Evaluation.

Undergraduate Student

Name: Mierzejewski, Timothy

Worked for more than 160 Hours: No

Contribution to Project:

Tim Mierzejewski is a CMU undergraduate student who worked with our research group as a research assistant in the Spring-2008 semester.
Tim assisted the graduate students with minor programming tasks.

Technician, Programmer

Name: Peterson, Erik

Worked for more than 160 Hours: Yes

Contribution to Project:

Erik Peterson is the main developer of the major components of our
transfer-based Machine Translation approach. He develops the statistical transfer engine (XFER engine) and the decoder for our MT research system, and integrates language-specific components for all language-pair specific prototype systems that we develop under LETRAS. Erik Peterson left the project at the end of May 2008.

Name: Vega, Rodolfo

Worked for more than 160 Hours: Yes

Contribution to Project:

Rodolfo Vega is our project coordinator for collaborations with research groups in South and Latin America. His contributions have been primarily in coordinating and assisting us in collaborations with Chile and exploring possible collaborations in Peru and Bolivia.

Other Participant

Research Experience for Undergraduates

Name: Andrews, David

Worked for more than 160 Hours: No

Contribution to Project:

David Andrews is a CS undergraduate student who has been working as a part-time research assistant on the project, helping us with code implementation for the new Joshua decoder. Supported by REU funds.

Name: Degolia, Joseph

Worked for more than 160 Hours: No

Contribution to Project:

Joseph Degolia is an undergraduate student, working as a part-time research assistant with Lori Levin on language technologies for Inupiaq. Supported by REU funds.

Name: Venkateswaran, Sai

Worked for more than 160 Hours: No

Contribution to Project:

Sai Prasnath Venkateswaran is an undergraduate student, working as a part-time research assistant with Lori Levin on language technologies for Inupiaq. Supported by REU funds.

Name: Mayer, Ida

Worked for more than 160 Hours: No

Contribution to Project:

Ida Mayer is an undergraduate student, working as a part-time research assistant with Lori Levin on language technologies for Inupiaq. Supported by REU funds.

Organizational Partners

University of Alaska Fairbanks Campus

Lawrance Kaplan and the Alaska Native Language Center at the University of Alaska, Fairbanks, are collobartive research partners with us on LETRAS on applying our MT approach to the development of MT technology between Inupiaq and English.

University of Haifa

Dr Shuly Wintner and his research group at the Computer Science Department at the University of Haifa, Israel, are collaborative research partners with us on applying our MT approach towards constructing MT technology between Hebrew and English.

University of Sao Paulo

Dr Marcello Modesto from the Linguistics Department at the University of Sao Paulo, Brazil, is a collaborative research partner, working with us on applying our MT approach to the development of MT Technology between Brazilian Portuguese and English, and between native languages of Brazil and Portuguese.

Sabanci University

Dr Kemal Oflazer and his students are collaborative research
partners with us on applying our Stat-XFER MT approach developed under LETRAS towards constructing MT technology between Turkish and English.

Dublin City University

We collaborate on research on advanced syntax-based statistical MT with DR. Andy Way and his research group. John Tinsley - a PhD student at DCU visited us at CMU in January 2009 and spend a month working closely with our group.

Other Collaborators or Contacts

Rodolfo Vega, who is affliated with our AVENUE and LETRAS projects, visited Bolivia in summer 2006 to explore possible research collaboration opportunities.

Rodolfo Vega and Carlos Fasola (graduate student at Rutgers University) visited Chile in early 2010 for meetings and data collection with representatives of the indigenous Mapuche population.

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Major research activity at CMU under the LETRAS grant was completed by end of July-2009 and was described in last year's annual report (attached). Collaboration activities with our external partners has continued over the past year, partly supported by supplements to the LETRAS grant. These primarily include work with the Alaska Native Language Center at the University of Alaska, Fairbanks (sub-contract under LETRAS); work on Mapuche with collaborators in Chile (LETRAS supplement); and work with University of Haifa, Israel on Hebrew-to-English and Hebrew-to-Arabic MT (LETRAS travel supplement).

Findings: (See PDF version submitted by PI at the end of the report)

The collaborations with Alaska and Israel have resulted in two new publications (see publication list). Findings from previous years are reported in the attached file.

Training and Development:

The project has partially or fully or supported five PhD students.

The project has also provided technical training and support for four external collaborating research groups, which use the MT software developed in the course of the project.

Outreach Activities:

The project has extensive international research collaboration activities. Discussions regarding potential new partnership and collaboration activities have involved senior governmental levels in Chile, Bolivia and Peru.

Journal Publications

Font Llitjos, A. and S. Vogel, "A Walk on the Other Side: Adding Statistical Components to a Transfer-Based Translation System", In Proceedings of Workshop on Syntax and Structure in Statistical Translation (SSST) at HLT-NAACL 2007, p. , vol. , (2007). Published,

Font Llitjos, A., J. Carbonell and A. Lavie, "Improving Transfer-Based MT Systems with Automatic Refinements", in Proceedings of MT Summit XI, Copenhagen, Denmark, p. , vol. , (2007). Published,

Lavie, A., "Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation", Invited paper in Proceedings of CICLing-2008, p. 362, vol. LNCS 49, (2008). Published,

Hanneman G, E. Huber, A. Agarwal, V. Ambati, A. Parlikar, E. Peterson, A. Lavie, "Statistical Transfer Systems for French-English and German-English Machine Translation", in Proceedings of the Third Workshop on Statistical Machine Translation at ACL-2008, p. 163, vol. , (2008). Published,

Lavie A., A. Parlikar, V. Ambati, "Syntax-Driven Learning of Sub-Sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora", Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2), p. 87, vol. , (2008). Published,

Ambati, V. and A. Lavie, "Improving Syntax-Driven Translation Models by Re-structuring Divergent and Non-isomorphic Parse Tree Structures", In Proceedings of Student Research Workshop at Conference of the Association for Machine Translation in the Americas (AMTA-2008)., p. 1, vol. , (2008). Published,

Hanneman, G., V. Ambati, J. H. Clark, A. Parlikar and A. Lavie., "An Improved Statistical Transfer System for French-English Machine Translation.", In Proceedings of the Fourth Workshop on Statistical Machine Translation at the 2009 Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2009)., p. 1, vol. , (2009). Published,

Hanneman, G. and A. Lavie, "Decoding with Syntactic and Non-Syntactic Phrases in a Syntax-Based Machine Translation System.", In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation at the 2009 Meeting of the North-American Chapter of the Association for Computational Linguistics (NAACL-HLT-2009)., p. 1, vol. , (2009). Published,

Ambati, V., A. Lavie and J. Carbonell., "Extraction of Syntactic Translation Models from Parallel Data using Syntax from Source and Target Languages.", In Proceedings of MT Summit XII., p. 1, vol. , (2009). Accepted,

Bills, Aric, Lori S. Levin, Lawrence D. Kaplan, and Edna Ahgeak MacLean., "Finite-state Morphology for Inupiaq.", In 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, p. 19, vol. , (2010). Published,

Reshef Shilon, Nizar Habash, Alon Lavie and Shuly Wintner, "Machine Translation between Hebrew and Arabic: Needs, Challenges and Preliminary Solutions", To appear in Student Research Workshop at AMTA-2010, p. 1, vol. , (2010). Accepted,

Books or Other One-time Publications

Lavie, A., E. Peterson, S. Wintner, D. Shacham and N. Melnik, "Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System", (2007). , Non-archival conference
Bibliography: Presented at BISFAI 2007 Conference, Ramat Gan, Israel

Font Llitjos, A. and W. Ridmann, "The Inner Works of an Automatic Rule Refiner for Machine Translation", (2007). , Published
Bibliography: Presented at METIS-II Workshop, Leuven, Belgium.

Lavie, A., "Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation", (2008). Book, Published
Editor(s): Alexander Gelbukh
Collection: Computational Linguistics and Intelligent Text Processing
Bibliography: Springer LNCS 4919

Ariadna Font Llitjos, "Automatic Improvement of Machine Translation Systems
", (2007). Thesis, Published
Bibliography: CMU-LTI-07-008

Web/Internet Site

URL(s):

http://www.lti.cs.cmu.edu/Research/Thesis/2007-AriadnaFontLlitjos.pdf

Description:
Ariadna Font-Llitjos PhD Dissertation

Other Specific Products

Contributions

Contributions within Discipline:
Our framework and approach to building MT systems is unique in the field, and combines the strengths of strengths of statistical approaches to MT, based on modern machine learning techniques, with deeper linguisticly motivated grammar-based methods. This general approach is attracting broad attention within the MT community over the last couple of years.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:
Our 20,000 word (3,100-sentence) 'Elicitation Corpus' was delivered to LDC, and is an integral part of 'LCTL language packs' being created at LDC under separate DoD funding. [LCTL = Less Commonly Taught Languages].

Contributions Beyond Science and Engineering:

Conference Proceedings

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Any Product
Contributions: To Any Other Disciplines
Contributions: To Any Human Resource Development
Contributions: To Any Beyond Science and Engineering
Any Conference

LETRAS Project - Activities: Year-3

The primary scientific goal of LETRAS is to explore novel methods for developing Machine Translation technology for languages for which only limited amounts of data and language resources are available. LETRAS is a follow-on project to the AVENUE project (IIS-0121631, 2001-2007), which was funded by an NSF/ITR grant. The primary underlying framework for our MT approach was developed in the course of AVENUE, and is based on a syntactic-transfer formalism, consisting of synchronous context-free rules that are augmented with feature unification constraints. The focus in the LETRAS project is on further development of the underlying general MT framework and expanding its application to new languages, including Inupiaq (a native Alaskan language), and native languages in Brazil and other countries in Latin America. We are also applying our MT methodology to the development of a Hebrew-to-English MT system, and collaborate with a group that is developing a Turkish-to-English MT system using our MT framework. We collaborate extensively with a number of other research groups on the development of MT research prototypes for the above mentioned languages: University of Alaska, Fairbanks (for Inupiaq-English MT), University of Sao Paulo, Brazil (for Brazilian Portuguese to English MT); with the University of Haifa, Israel (for Hebrew-to-English MT); and with Sabanci University in Turkey (for Turkish-to-English MT).

Over the past two years, our MT framework has been extensively extended and redesigned into a fully implemented syntax-based framework for building MT systems from large volumes of parallel data. We chose the name "Statistical Transfer" for this new framework, or in short - "Stat-XFER". The Stat-XFER framework has been used to develop large-scale MT systems for Arabic-to-English (for the NIST-09 MT Evaluation), Chinese-to-English (under funding from the DARPA GALE Program), for French-to-English (for the WMT-08 and WMT-09 shared task competitive evaluation), and German-to-English (WMT-08 only). We also developed a Stat-XFER system for Urdu-to-English under limited resources, as part of the NIST-08 and NIST-09 MT Evaluations.

Our main research and education activities in the third year of the project have been the following:

(1) Continued research on the design and implementation of a framework for transfer-rule based Machine Translation, that can support compositional transfer-rule grammars that are either manually developed or automatically acquired from data. This includes the following main sub-activities:
- Incremental improvements to our Stat-XFER MT engine, to enable effective translation performance with large scale translation lexicons and grammars.
- Continued research into automatic acquisition of transfer rules from various types of parallel data, including both controlled "elicited" data that is manually word-aligned, as well as large-scale parallel corpora, where parallel sentences are automatically word-aligned.
- Continued research into automatic acquisition of synatax-based word and phrase translation lexicons from sentence-parallel corpora, under various different resource scenarios (parse trees on one side and/or both sides of the parallel corpus).
- New extensive research on methods for combining and joint-decoding with automatically-extracted syntactic and non-syntactic phrases.

(2) Development of MT prototype systems using the Stat-XFER MT framework. Several large scalse MT prototype systems were developed over the past three years exclusively by our group at CMU. Additional MT prototype

systems were developed in collaboration with our research partners (see more below). The prototype systems developed exclusively at CMU include:

- Urdu-to-English: We developed an Urdu-to-English Stat-XFER system under the low-resource "Urdu Track" that was included in the 2008 and 2009 NIST MT Evaluations. The system was developed using the limited resources released by LDC and NIST for this track, using our Stat-XFER methods and using the open-source Moses framework. The system was entered for evaluation in both NIST MT-08 and MT-09.
- Arabic-to-English: We developed an Arabic-to-English Stat-XFER system that incorporated our latest work on combining syntactic and non-syntactic resources, and submitted the system for evaluation in NIST MT-09.
- French-to-English and German-to-English: We developed prototype MT systems for both language pairs under the shared task that was part of the WMT-08 Statistical MT workshop (held at ACL-2008). The systems were develoepd on Europarl parallel data that was released by WMT-08. The systems were entrered for evaluation to WMT-08. An updated version of the French-to-English Stat-XFER system, incorporating our latest methods for combining syntactic and non-syntactic phrases was developed and submitted to WMT-09.
- Chinese-to-English: We developed a full-scale Stat-XFER system for Chinese-to-English translation under a small amount of funding provided by the DARPA GALE program, as part of the IBM-led Rosetta consortium. The Chinese-to-English Stat-XFER system was one of several different MT systems that were combined together for the GALE phase-2 evaluations.

(3) Collaborative research work with our partners in Alaska, Brazil, Israel, and Turkey. Our main collaborative activities included:

- Alaska: We collaborate with Dr. Lawrence Kaplan and Dr. Edna McLean from the Alaska Native Language Center at the University of Alaska, Fairbanks, on the application of our MT technology to MT for Inupiaq. Activities in the past year included a research workshop, that was held in April 2009.
- Brazil: We collaborate with Dr. Marcello Modesto from the Linguistics department at the University of Sao Paulo (USP) in Brazil on developing a Brazilian Portuguese to English prototype MT system under our Stat-XFER MT framework. An MS student at USP is doing her MS thesis work under this collaborative project.
- Israel: We collaborate with Dr. Shuly Wintner and his research group at the University of Haifa in Israel. We jointly develop a prototype MT system for Hebrew-to-English based on our Stat-XFER MT framework.
- Turkey: We collaborate with Dr. Kemal Oflazer and his research group at Sabanci University in Turkey. We provided technical support to Dr. Oflazer and his student for developing a prototype MT system for Turkish-to-English using our Stat-XFER MT framework.

LETRAS Project - Findings: Year-3

Major findings and accomplishments in the third year of the project:

(1) Improvements to our core MT engine, to enable effective translation performance with large scale translation lexicons and grammars.
(2) Improvements to our monotonic "decoder" that works in tandem with the transfer engine in order to produce complete sentence translations. The decoder can "jointly-decode" by combining syntactic fragments put together by our transfer engine with "non-syntactic" phrases extracted using the state-of-the-art Moses framework.
(3) Continued research into automatic acquisition of transfer rules from various types of parallel data. The main focus this year was on acquisition from large-scale parallel corpora, where parallel sentences are automatically word-aligned. New methods for simultaneously improving word alignments and phrase extraction are being explored.
(4) Continued research into automatic acquisition of synatax-based word and phrase translation lexicons from sentence-parallel corpora, under various different resource scenarios (parse trees on one side and/or both sides of the parallel corpus).
(5) New research into methods for effectively combining automatically extracted syntax-based translation lexicons with non-syntactic phrases extracted using traditional SMT heuristic methods.
(6) Participation in MT Marathon in Prague: In January 2009, co-PI Lavie and two students (Greg Hanneman and Jonathan Clark) participated in the MT Marathon workshop in Prague. As part of the workshop, we performed an intensive one-week group exercise of constructing a prototype Stat-XFER MT system for Czech-to-English.
(7) Brazil: We collaborate with Dr. Marcello Modesto from the Linguistics department at the University of Sao Paulo in Brazil. Dr. Modesto visited CMU for three weeks in summer 2006 and for an additional two weeks in March 2007, during which an initial Portuguese-to-English MT system was constructed. co-PI Dr. Alon Lavie visited Sao Paulo Brazil for one week in May 2008, to conduct joint research work with Dr. Modesto and Lucia da Silva - an MS student. Dr. Lavie also gave an invited lecture at the University of Sao Paulo on Machine Translation and our MT framework. Lucia da Silva, the MS student from Brazil, visited CMU in February 2009 to continue research work on the system.
(8) Israel: We collaborate with Dr. Shuly Wintner and his research group at the University of Haifa in Israel. We jointly develop a prototype MT system for Hebrew-to-English based on our AVENUE/LETRAS MT framework. Dr. Wintner has spent the 2006/20077 academic year on sabbatical at CMU, primarily working on another NSF-funded project, but his visit has also facilitated progress on our MT project. co-PI Alon Lavie made two one-week visits to Israel during 2007, participated in two workshops where the project was highlighted, and gave a presentation about the project at a high-visibility local conference. Dr. Lavie also was an invited keynote speaker at the CICLing-2008 conference in Haifa, Israel in February 2008. The invited presentation focused on the Stat-XFER MT framework. Dr. Lavie visited Israel in April 2009 and met with Dr. Wintner and his students.
(9) Alaska: A two-day workshop was held in April 2009 to assess progress and plan future research activities for Inupiaq. External particpants included Professor Lawrence Kaplan (Director of the Alaska Native Language Center) and Dr. Edna MacLean (Inupiaq native speaker and language specialist) from the University of Alaska, Fairbanks.