The Linguistic-Core Approach
to Structured Translation and Analysis of Low-Resource Languages

2011 Program Review for ARL MURI Project

4 November 2011

The Cast

CMU:

Jaime Carbonell 

Lori Levin 

Stephan Vogel 

Noah Smith 

ISI:

Kevin Knight 

David Chiang 

MIT:

Regina Barzilay 

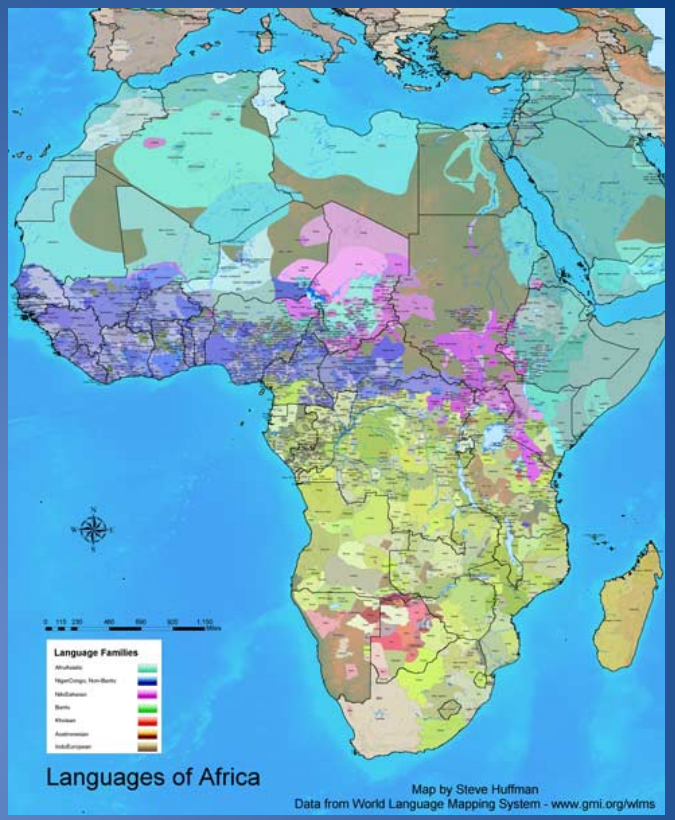
UT:

Jason Bladridge 

Supporting roles: **8** Graduate Students, **2** Postdocs, **N** Informants, …

The Plot

- Objectives
  - MT & Analysis
  - Of Low-resource Languages
  - Based on Linguistics
  - Supported by Stat Learning
- Challenges
  - Data sparsity & collection
  - Major Divergences from Eng
  - “Universal” Solutions (vs just for a given language)



- Selected Languages
  - Kinyarwanda
    - Bantu (7.5M speakers)
  - Malagasy
    - Malayo-Polynesian (14.5M speakers)

The Setting (from Proposal)

- LR Languages, e.g. in Africa, cannot be ignored
- MT & TA for LRLs requires a linguistic core
  - Insufficient parallel text for standard SMT
  - Insufficient annotations for purely statistical TA
- Phrasal SMT, even for HRL, errs
  - E.g. divergences, long-distance movements,…
- But Computational Linguists are Expensive
  - Cannot dedicate person-centuries per language to write, test and debug massive rule-based systems
  - Army needs a more rapid & cost effective approach

The Scientific Questions

- Can deep linguistic representations benefit practical MT & TA?
- Can we marry learning from data with expert-crafted declarative linguistics?
- Can we uncover underlying linguistic structure through comparative language analysis?
- How can we extend MT-motivated linguistic-core capabilities to related TA tasks?
- Can different linguistic analyses reinforce each other synergistically?
- How important is resolving complex morphology?
- How important are general semantic features for MT?
- How well can unsupervised learning methods augment linguistically motivated analyses for MT and TA?
- ….

Act I: Exploratory Research

Scene I: Data

- Obtained Malagasy Bible and align with modern English Bible.
- Converted 33 KGMC multilingual transcripts (Kinyarwanda, English, French) of interviews of survivors of the Rwandan genocide to clean, aligned XML.
- Created seed datasets for Malagasy and Kinyarwanda from the linguistic literature and annotate them for syntactic structure.
- Reached out to Rwandan and Malagasy communities to find native speakers,
- Translate three BBC Rwanda articles to English and annotate.
- Translated Malagasy website articles (Lakroa and Lagazette) to English and annotate with syntactic structures
- Adapted Malagasy morphological transducer from Dalrymple et al and annotate several sentences based on its output.
- Annotated KGMC transcripts for syntactic structure (about 100 trees).
- Created tools for supporting consistent annotations that will work for MT researchers (tokenizers, tree validators).
- Crowd-sourcing for non-linguist native-speaker data collection.
- Active learning for focusing on most valuable missing data.
- Curated data releases 1.0 and 2.0 for Malagasy, Kinyarwanda, and English.

Act I: Exploratory Research

Scene 2: Linguistic Core + ML

- Rule-based Kinyarwanda morphological analyzer.
- Development of semantic representation graphs for general-purpose translation.
- Development of probabilistic acceptors and transducers for graph structures.
- Tokenizer for Kinyarwanda and Malagasy.
- Completed formalism design (dependency to dependency MT)
- Investigate hand-written synchronous tree-adjoining grammar rules for Kinyarwanda .
- Learning syntactic structure from sparse semantic representations.
- Learning unsupervised morphology by modeling syntactic context.
- Upparse unsupervised parsing methodology based on finite-state methods and evaluated on English, German and Chinese data.
- Bilingual part of speech model based on feature-rich Markov random fields.
- Method for transferring information in supervised models for one or more resource-rich languages to an unsupervised learner for a resource-poor language, tested on part-of-speech tagging and parsin.
- Model for discovering multi-word, gappy expressions in monolingual and bilingual text, evaluated within a translation system
- Model for word alignment based on feature-rich conditional random fields

Act I: Exploratory Research

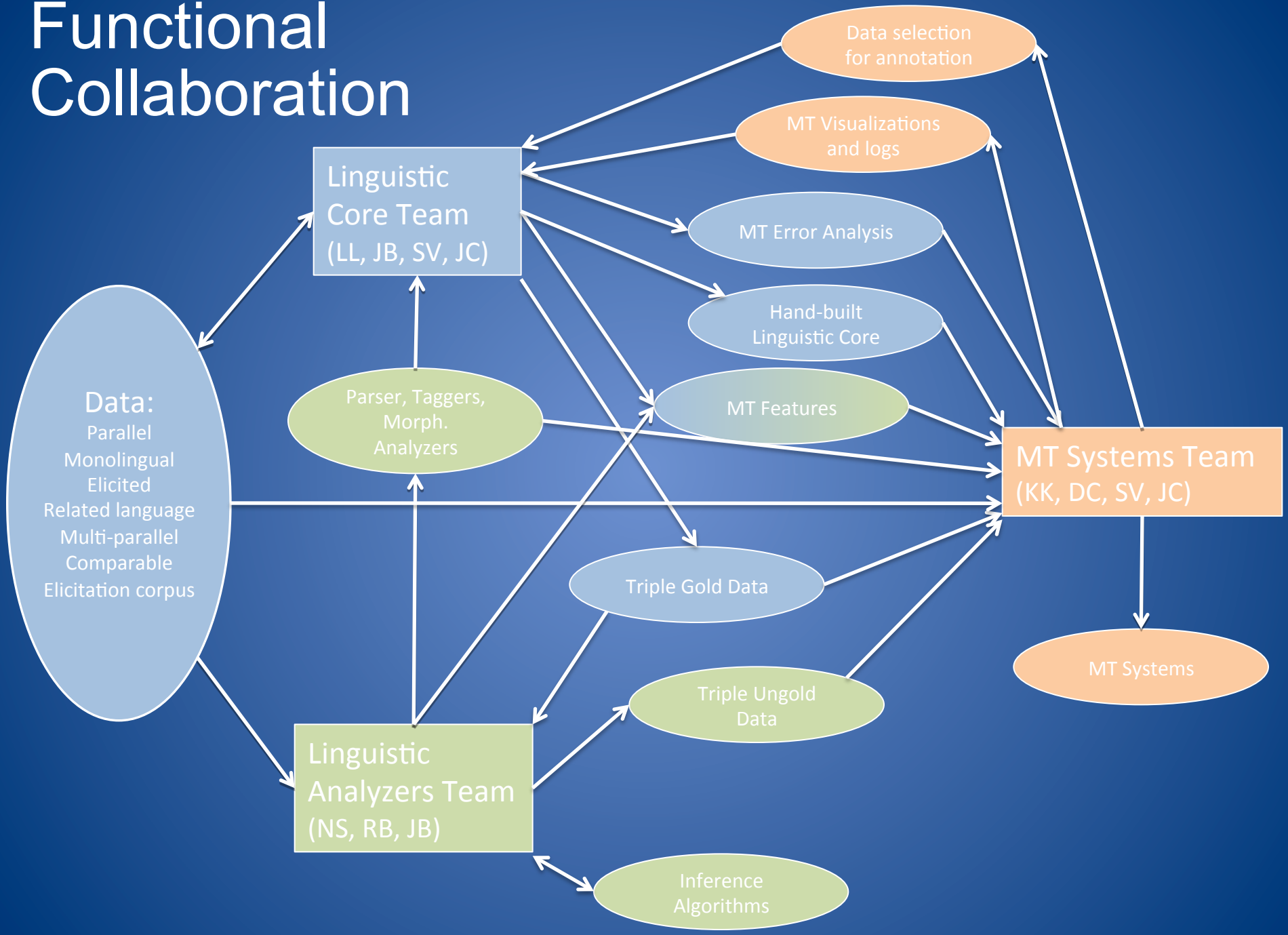
Scene 3: MT Frameworks & Systems

- Phrase-based Malagasy and Kinyarwanda systems build for initial baseline
- Four end-to-end MT systems (m2e and k2e, using Hiero and syntax-based MT systems).
- Kinyarwanda Synchronous-grammar (SAMT) system build
- Implemented a hierarchical phrase-based German-English translation system that was ranked #2 in the SMT competition (after Google). This system incorporated a discriminative German parsing model
- Designed, implemented, and tested a new translation model based on dependencies over phrases.
- Explored methodology for testing hypotheses about translation systems, leading to practical recommendations for researchers in the field.
- Translation systems targeting Kinyarwanda and Malagasy incorporating the data developed by MURI collaborators were developed, and improvements using CRF word alignments were replicated in these new language pairs.

Our Slightly-Revised Approach

- Linguistic core: Universals & Specifics
  - Specialize core to each language pair minimally as needed
  - Active Learning when annotations/translations required
  - Unsupervised learning when possible
- Parallel activities:
  - Development of training/testing data & annotations (on targeted L's)
  - Linguistic analysis (of targeted L's)
  - Core linguistic engine development (on other L's)
- Exploration of multiple paradigms
  - E.g. Dependency parsing
  - E.g. Finite-state transducers
  - Ensemble methods
- Build, evaluate, refine glass-box end-to-end prototypes
  - Requires baselines, and end-to-end MT systems

Functional Collaboration



Publications

- Vamshi Ambati, Stephan Vogel and Jaime Carbonell. "Collaborative Workflow for Crowdsourcing Translation", To Appear in the 2012 ACM Conference on Computer Supported Cooperative Work, Washington, USA
- Vamshi Ambati, Stephan Vogel and Jaime Carbonell. "Multi-Strategy Approaches to Active Learning for Statistical Machine Translation", Accepted to the 13th Machine Translation Summit, Xiamen, China, 2011
- Vamshi Ambati, Stephan Vogel and Jaime Carbonell. "Towards Task Recommendation in Micro-Task Markets" , In the Proc. of the 3rd workshop on Human Computation, AAAI. 2011.
- Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel and Jaime Carbonell. "Active Learning with Multiple Annotations for Comparable Data Classification Task", In the Proc. of Building Comparable Corpora Workshop, ACL. 2011.
- Desai Chen, Chris Dyer, Shay B. Cohen, and Noah A. Smith. Unsupervised Bilingual POS Tagging with Markov Random Fields. In Proceedings of the First Workshop on Unsupervised Learning in NLP. 2011.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better Hypothesis Testing for Machine Translation: Controlling for Optimizer Instability. In Proc. of ACL. 2011.
- Shay B. Cohen, Dipanjan Das, and Noah A. Smith. Unsupervised Structure Prediction with Non-Parallel Multilingual Guidance. In Proc. EMNLP. 2011.

More Publications

- Chris Dyer, Kevin Gimpel, Jonathan H. Clark, and Noah A. Smith. The CMU-ARK German-English Translation System. In Proc. WMT. 2011.
- Chris Dyer, Jonathan H. Clark, Alon Lavie, and Noah A. Smith. Unsupervised Word Alignment with Arbitrary Features. In Proceedings of ACL. 2011.
- Kevin Gimpel and Noah A. Smith. Quasi-Synchronous Phrase Dependency Grammars for Machine Translation. In Proc. EMNLP. 2011.
- Kevin Gimpel and Noah A. Smith. Generative Models of Monolingual and Bilingual Gappy Patterns. In Proc. WMT. 2011.
- Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings ACL. 2011.
- Yoong Keok Lee, Aria Haghighi and Regina Barzilay. Modeling Syntactic Context Improves Morphological Segmentation. In Proc. of CoNLL, 2011.
- Tahira Naseem, Regina Barzilay. Using Semantic Cues to Learn Syntax. In Proc. AAAI 2011.
- Elias Ponvert, Jason Baldridge and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In Proceedings of ACL 2011.
- Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati and Stephan Vogel, “CMU Haitian Creole-English Translation System”, In Proc. WMT. 2011.

THANK YOU!