Human Language Technologies (HLT) Workshop 2006

Machine Translation and Statistical Language Learning in the KSU Lab for Knowledge Discovery in Databases

William H. Hsu
Joint work with: Waleed Al-Jandal, Tejaswi Pydimarri, Chris Meyer
Tuesday, 30 May 2006

Laboratory for Knowledge Discovery in Databases
Kansas State University
http://www.kddresearch.org/KSU/CIS/HLT-General-20060530.ppt

HLT Research at Kansas State: Scope, Goals and Technical Objectives
OUTLINE

- Background, Related Work and Rationale
- Technical Objectives and Significance
- Development Plan
- Preliminary Progress Report
- Future Directions: Opportunities for Collaboration

PROBLEM STATEMENT: MACHINE TRANSLATION

- Basic Task Specification
  - Source: foreign sentence \(f\)
  - Target: native sentence \(e\) (e.g., English)
- Input
  - Parallel training corpora (documents) or speech: \((f, e)\) pairs
  - Usually, but not always, produced manually
- Expected Output: translations \(e\) for new sentences \(f\)
- What Does This Mean?
  - Alignment, parsing tasks
  - Interactive, possibly real-time, translation tasks
Overview: Enabling Technologies

- Artificial Intelligence
- Bayesian Methods
- Cognitive Science
- Communications
- Computational Theory
- Communications
- Linguistics
- Neuroscience
- Psychology
- Statistics

Intelligent Natural Language Processing

Overview:

- Gisting
- Publication
- Speech
- Complexity Bounds
- Learnability
- Maximum Entropy
- MDL Approaches
- Optimal Codes
- Language Learning
- Learning to Reason
- Estimation
- Discriminative Models
- Generative Models
- EM
- Analysis of Variance
- Evaluation
- Nonparametric Inference
- Human Performance
- Knowledge Acquisition
- Judgment & Decision Making
- Memory
- Cognitive Neuropsych
- Syntax
- Morphology
- Hierarchies
- Analysis of Variance
- Evaluation
- Nonparametric Inference
- Human Performance
- Knowledge Acquisition
- Judgment & Decision Making
- Memory
- Cognitive Neuropsych
- Syntax
- Morphology
- Hierarchies

Background [1]: Translation Strategies

- MT Strategies (1954-2006)
  - Electronic dictionaries
  - Hand-built by experts
  - Hand-built by non-experts
  - Original direct approach
  - Typical transfer system
  - Classic interlingual system

- Shallow/Simple
  - Word-based only
  - Phrase tables
  - Learn from annotated data
  - Original statistical MT
  - Example-based MT
  - Learn from un-annotated data
  - Fully automated

- Deep/Complex
  - Knowledge Representation Strategy
  - Syntactic
  - Constituent Structure
  - Semantic analysis
  - Focus of New Research

- All manual
- Knowledge Acquisition Strategy
- All manual

Slide courtesy of Laurie Gerber
insistent Wednesday may
recurred her trips to Libya
tomorrow for flying
Cairo 6-4 ( AFP ) - an official
announced today in the Egyptian
lines company for flying " Tuesday
is a company “insistent for
flying”, may resumed a
consideration of a day Wednesday
tomorrow her trips to Libya of
Security Council decision trace
international the imposed ban
comment.
And said the official “ the
institutions sent a speech to
Ministry of Foreign Affairs of
lifting on Libya air, a situation her
receiving replying are so a trip will
pull to Libya a morning
Wednesday “.

Egyptair Has Tomorrow to
Resume Its Flights to Libya
Cairo 4-6 (AFP) - said an official at
the Egyptian Aviation Company
today that the company egypair
may resume as of tomorrow,
Wednesday its flights to Libya
after the International Security
Council resolution to the
suspension of the embargo
imposed on Libya.
“ The official said that the
company had sent a letter to the
Ministry of Foreign Affairs,
information on the lifting of the air
embargo on Libya, where it had
received a response, the first take
off a trip to Libya on Wednesday
morning “.

Related Work:
Phrase-Based Learning for SMT

1. phrase alignments from word-aligned model
   * Used in GIZA++ toolkit [Och & Ney, 2000]
   * See IBM models [Brown, 1993]
2. linguistically motivated models
   * [Yamada & Knight, 2001; Imamura, 2002]
   * Require subtree matching in syntax tree (parse tree)
3. joint phrase model
   * [Marcu & Wong, 2002]
   * Directly learns phrase-level alignment of parallel corpus
4. generative phrase alignment [Koehn, Och & Marcu, 2003]
5. hierarchical models [Chiang, 2005; Taskar, 2005]
Rationale

- **What Works:** Phrase-Based Translation Methods
- **Who it Works for & When & Where it Works**
  - 5S: Streams, Structures, Spaces, Scenarios, Societies
  - Application context: the “performance element”
- **How/Why it Works:** Synthesis of Technologies
  - 1. Moore’s Law: Advances in Processing Power
  - 2. Better Metrics: Bilingual Evaluation Understudy (BLEU)
  - 3. Bigger Corpora: Arabic, Chinese

Outline

- Background, Related Work and Rationale
- **Technical Objectives and Significance**
- Development Plan
- Preliminary Progress Report
- Future Directions: Opportunities for Collaboration
Limitations of Current State of the Art

- Applications: How Far Can Current Methods Take Us?
- Role of Knowledge
  - Correction models
  - Context-specificity
- “Use What Works”: Brute-Force Technology?
  - Unsatisfying for grammarians, semantics researchers
  - Some successes with latent semantic analysis in NLP
- Info Retrieval (IR) vs. Extraction (IE), Understanding
- Utility Measures: Are Metrics Meaningful?

Novel Contributions [1]: Context-Specific Machine Learning

- Using Context
  - Word-sense disambiguation (Roth, 1998)
    - Homonyms
    - Part-of-speech tagging
  - Context-Specific Independence
    - Knowledge maps aka probabilistic similarity networks (Heckerman, 1991)
    - Graphical models (Boutilier et al., 1996)
    - Entity clustering (Barash & Friedman, 2002)
- Detection of Hidden Changes in Context
- Contextual Correction
Novel Contributions [2]:

Relational Knowledge Representation

- First-order relational models
  - Description logics
  - Graphical models
- Representation: Bridging Learning and Reasoning
- Semantics
  - Traditional wisdom: tradeoff
  - New idea (cf. Koller, 2001)
    - Greater expressiveness
    - Lower complexity

Novel Contributions [3]:

Learning in Graphical Models

Continuing Work:
- Speeding up Approximate Inference using Edge Deletion - J. Thornton (2005)
- Bayesian Network tools in Java (BNJ) v4 - W. Hsu, J. M. Barber, J. Thornton (2006)
Technical Plan:
Development Objectives, Progress, and Collaboration Opportunities

Outline

- Background, Related Work and Rationale
- Technical Objectives and Significance
- Development Plan
- Preliminary Progress Report
- Future Directions: Opportunities for Collaboration
Development Plan: Approximate Timeline

- **2005**
  - Spring: Statistical Machine Translation (SMT) Group founded
  - Fall: SMT seminar – 12 papers on state of the field
  - Resource: SMT bibliography
- **2006**
  - Spring: getting set up with corpora, GIZA, BLEU
  - Summer: NIST evaluation; BNJ v4 release
  - Fall: SMT seminar – lessons learned; Targeted Excellence on HLT
  - Winter: how-to workshop on MT techniques; BNJ v5 development
- **2007 and Beyond**
  - Spring: regional workshops on Learning, Memory, Cognition
  - Fall: tools integration

Bayesian Network tools in Java (BNJ) v4

© 2005 KSU Bayesian Network tools in Java (BNJ) Development Team
OUTLINE

- Background, Related Work and Rationale
- Technical Objectives and Significance
- Development Plan
- Preliminary Progress Report
- Future Directions: Opportunities for Collaboration

PRELIMINARY PROGRESS REPORT:
2001-2006

- 2001: Start of KSU Bioinformatics program
- 2002: EPSCoR First Award, bioinformatics, BNJ v2
- 2003: Summer KSU REU in Bioinformatics, BNJ v2, SRL-2003
- 2004: NSF ITR & FIBR; BNJ v3; ICSNW workshop, PODS-2004
- 2005: Start of Statistical Machine Translation (SMT) Group
- 2006: Learning, Memory, and Cognition Working Group
  - Development of end-to-end SMT system inspired by GIZA
  - Registration for 2006 NIST Evaluation
  - BNJ v4 & 5
  - HLT Targeted Excellence proposal
- 2007 & Beyond: interfaces – new BNJ/GEM, WEKA, ECJ
Work in Progress

- **End-to-end Statistical Machine Translation System**
  - Flexible, modular tools for
    - Alignment
    - Parsing
    - Phrase-based learning
    - Transformation-based learning (cf. Brill)
  - New modules substituted into infrastructure as completed

- **Comparisons with New Corpora**
  - Media studies: political journalism, large-volume data mining
  - Language studies: historical linguistics, etc.

- **New Metric Development**

Methodology

- **Overall Scientific Approach**
  - Using context-specific learning
  - Classification-based error detection
    - Committee machines: bagging & boosting
    - Mixture models: hierarchical mixture of experts (HME), etc.
    - Cascade filters
  - Integrative semisupervised learning
    - Clustering
    - Human categorization and ontology development

- **Applications-Oriented: Real Translation Tasks**
- **User-Centric: Real Task-Specific Metrics**
Next Steps

- Establishing an Interdisciplinary Research Initiative
  - K-State / KU / UNL collaboration
  - Resources: Linguistic Data Consortium
  - NIST evaluations

- Involving End Users of Machine Translation
  - Document users
  - Machine learning, data mining, info extraction researchers

- Novel Applications
  - Social networks and collaborative recommendation
  - Gisting and beyond

Outline

- Background, Related Work and Rationale
- Technical Objectives and Significance
- Development Plan
- Preliminary Progress Report
- Future Directions: Opportunities for Collaboration
Opportunities for Collaboration [1]:
Computational Sciences

- Information Extraction and Intelligent IR
  - Learning models for IE: ontologies
  - Latent semantic analysis
- Machine Learning
  - Natural language learning
  - Time series learning and understanding
  - Relational and first-order models
- Automated Reasoning
  - Probabilistic
  - Case-based and analogical
- Data Mining and Warehousing
- Grid Computing

Opportunities for Collaboration [2]:
Application and User-Centric Disciplines

- Anthropology
- Human Factors
- International Studies
  - Policy Studies
  - Business: Trade, Finance
  - Cultural Studies
- Journalism
- Library Science
- Modern Languages
- Political Science
Opportunities for Collaboration [3]: Linguistics and Psychology

- Cognitive Science
  - Intelligent systems and cognitive modeling
  - Cognitive neuropsychology: lesion studies, fMRI, etc.
- Educational Psych: Human Language Acquisition
- Ergonomics and Human Factors
- Linguistics: Computational Models of Language Production
- Judgment and Decision Making
  - Computational linguistic models of dialogue, negotiation
  - Utility-theoretic models of translation evaluation
- Psycholinguistics
  - Computational models
  - Translation as experimental test bed

Desiderata

- Usability (Q&A)
- Ergonomics
- Accessibility
- View control

Elements

- Unified data model
- Visualization widgets
- Figures of merit, evaluation mechanism (cf. BNJ)

Outreach: HCII Overlap & Tech Transfer
References [1]


References [2]

ACKNOWLEDGEMENTS

- KSU Psychology: Greg Monaco, Les Loschke
- KSU: Other Collaborators and Affiliates
  - Abdel Kader Kara, Talat Rahman, Dean Zollman, KSU Physics
  - Lori Bergen, KSU Journalism
- Abroad
  - Dan Roth, Cinda Heeren, Jiawei Han, AnHai Doan (USA, University of Illinois at Urbana-Champaign)
  - Violetta Cavalli-Svorza (USA, Carnegie Mellon University)
  - Susan Gauch (USA, University of Kansas)
  - Abdelhadi Soudi (Morocco)
  - Kirsten Hildrum (IBM T. J. Watson Labs)
  - More collaborations sought in HLT and HCII!

QUESTIONS AND DISCUSSION