# Machine Translation in TIDES Planning Committee Report

Kevin Knight, USC/ISI

Lori Levin, CMU

Young-Suk Lee, MIT-LL

Salim Roukos, IBM

Alex Waibel, CMU

# Technical Objectives

- Convert free text from a variety of languages into English.


- Foreign input:
  - text, audio transcripts, or summaries
- English output:
  - must be sufficiently accurate and fluent that a user can quickly and reliably assess relevance
  - should be extractable, detectable, summarizable

# Source Languages

- Primary:        Chinese, Arabic
  - focus for resource-intensive MT research

- Secondary:      Korean, Japanese
  - depending on availability of a million-word bilingual corpus

- Tertiary:        low-density languages
  - focus for targeted experiments

# Good News

- Lots of new ideas in data-driven MT research.
- Many more bilingual data resources online.
- Faster machines, larger memories.

# Bad News

- Bilingual data must be cleaned and sentence-aligned.
- As yet, no standard test suites or evaluations.

# Evaluation Metrics

- Good MT preserves meaning and is fluent.
    - traditionally difficult to quantify.

- Metrics should:
    - drive development of new technology
    - (vs. seeing what we can do with *old* technology)
    - allow for comparison of different systems.
    - demonstrate progress over several years.
    - encourage researchers to "hill climb" on a quantitative score.
    - send researchers back to the drawing board with specific problems to solve.

# Evaluation Metrics

- Fully automatic scoring
  - new system developed by IBM, called BLEU
  - answer key provided by multiple human translators
  - figures correlate with subjective quality assessments
  - no easy way to "beat the system"

- Human scoring
  - in use at several TIDES MT sites
  - establish correlations with automatic scoring

- Task-based scoring
  - to be done in TIDES IFEs

# Bilingual Evaluation Understudy:
# BLEU Metric vs. Human Judgements
## Native and Bilingual English speakers



Source: IBM
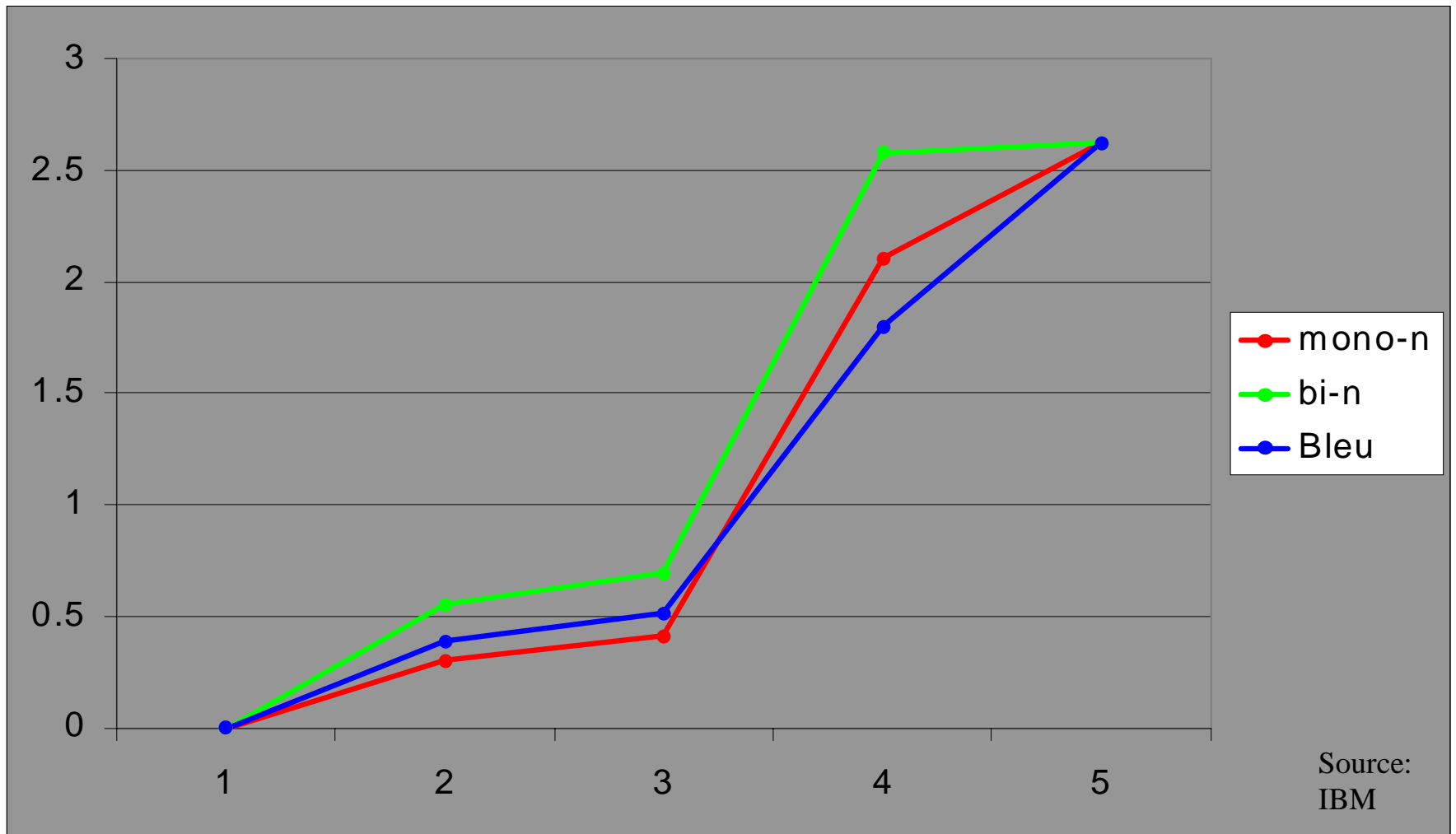
# Task Classes

- Heavyweight class
  - no limits
  - achieve best output quality possible

- Middleweight class
  - all sites use same bilingual corpus and dictionary
  - controlled comparison of models & methods

- Lightweight class
  - low-density languages
  - severe resource limitations

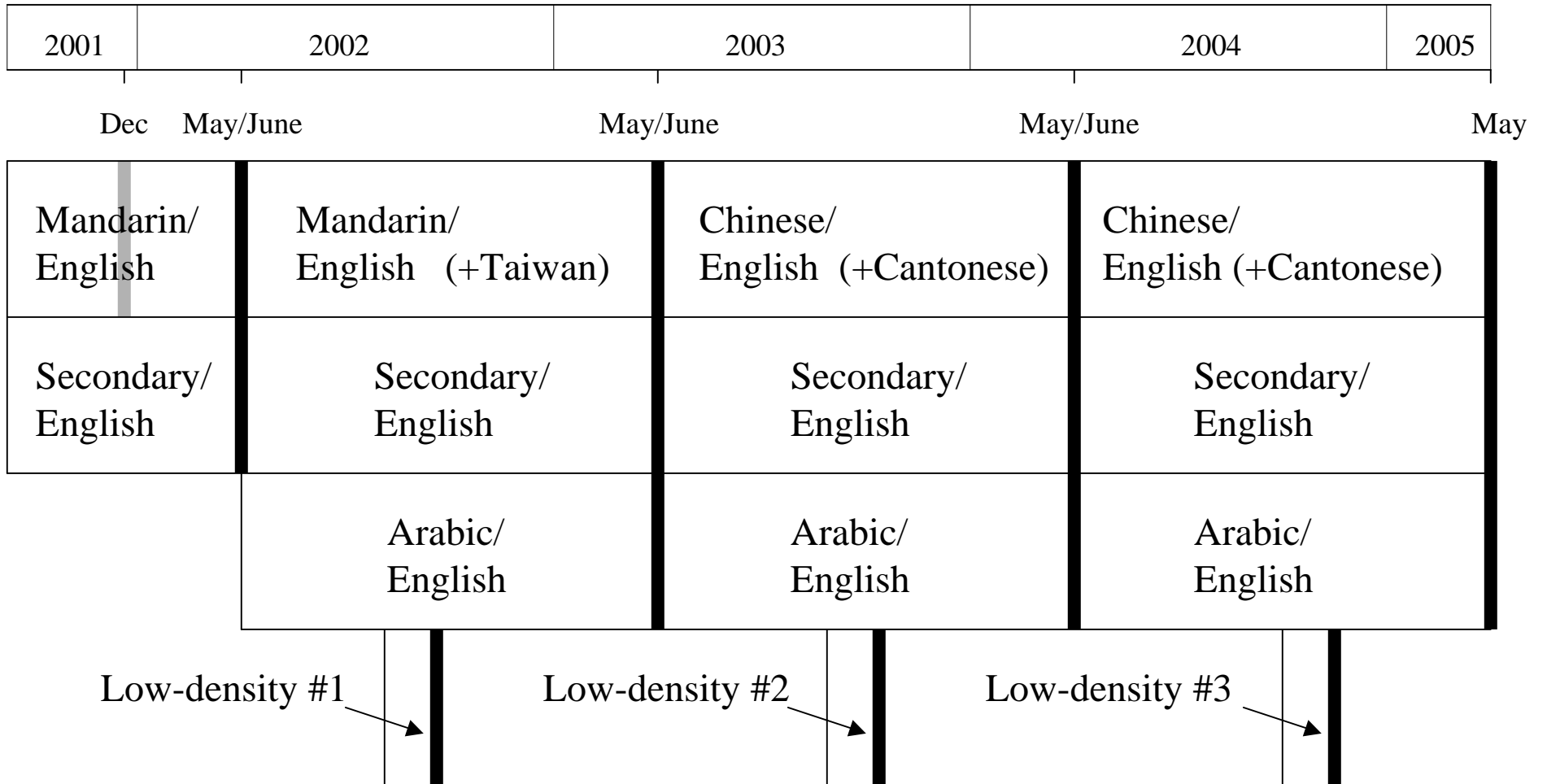# Test Material

- Test Genre:
  - foreign-language news reports

- Need to identify specific sample text streams of interest, e.g.:
  - www.chinanews.com.cn
  - www.eastday.com/epublish/gb/
  - www.cctv.com/news/china/
  - www.worldreport.com.cn/worldrep/
  - latelinenews.com/11/chinese/
  - search.peopledaily.com.cn/
  - www.joins.com (Jungang Daily, Korea)
  - www.chosun.com (Chosun Daily, Korea)

# Evaluation Schedule

| 2001 | 2002 | 2003 | 2004 | 2005 |
|------|------|------|------|------|
| Dec | May/June | May/June | May/June | May |

| Mandarin/ English | Mandarin/ English  (+Taiwan) | Chinese/ English  (+Cantonese) | Chinese/ English (+Cantonese) |
|---|---|---|---|
| Secondary/ English | Secondary/ English | Secondary/ English | Secondary/ English |
| | Arabic/ English | Arabic/ English | Arabic/ English |

Low-density #1

Low-density #2

Low-density #3

# Coordination With IDES

- To support integration across T, D, E, S:
  - Develop APIs for modules
  - Ensure overlap in linguistic resources developed for different modules

# Corpora and Other Resources

- Bilingual texts needed:
  - 10m words of bilingual news text (primary languages)
  - 1m words bilingual (secondary language)
  - 100k words bilingual (tertiary languages)
- Monolingual comparable texts needed:
  - 100m words (primary)
  - 10m words (secondary)
  - 1m words (tertiary)
- Other
  - Public, flexible sentence-alignment software
  - Central cleaning and alignment of all LDC parallel data
  - Bilingual lexicons (Mandarin, Korean, Arabic w/English)
  - Monolingual/parallel treebanks

# Technical Approaches in TIDES

**CMU (Rapdid)**

 example-based MT, statistical, multiple engine MT

**CMU (Nice)**

 transfer MT, learning from controlled elicitation corpus

**Cornell/CoGenTex/Montreal**

 transfer MT, annotated corpora

**IBM**

 statistical MT, maximum entropy, comparable corpora

**JHU/Maryland**

 statistical, projecting English tools onto other languages

**MIT Lincoln Laboratory**

 interlingua-based MT, prob. parsing, word sense disambig.

**USC-ISI (ReWrite)**

 statistical MT, syntax models, comparable corpora