

Scanned
pp 1-12

Translation Research: Overviews and Issues

Kevin Knight
USC/ISI

TIDES MT 2004

- HLT conference
 - May 2004
- MT common evaluation run by NIST
 - May 2004
- MT evaluation workshop
 - June 2004
- Integration/demonstration projects
 - Arabic TV translation (BBN, Virage)
 - eTIRR (Iraq Reconstruction Report)

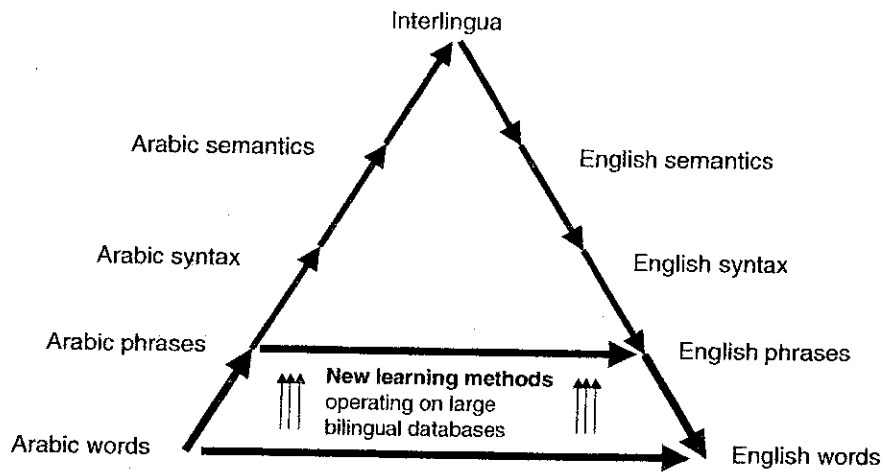
MT Data

- Thanks to Xiaoyi Ma and Mark Liberman at LDC!
- New parallel-text resources in 2004:
 - Chinese/English: 142m words → 210m words
 - Arabic/English: 74m words → 104m words
 - Arabic/English news: 600k words → 3m words

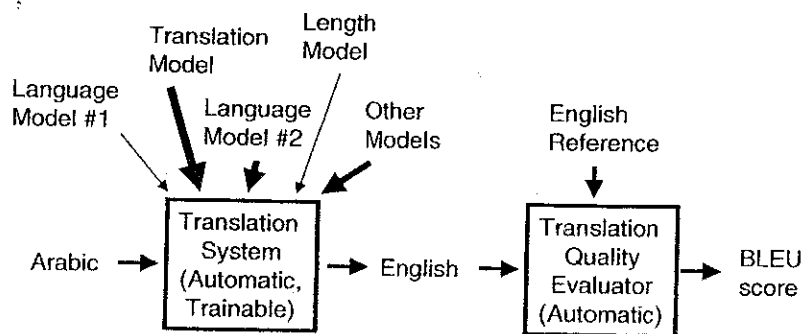
MT Evaluation

- BLEU score adopted by all sites
 - Invented at IBM
 - N-gram precision: overlap of MT output with multiple reference translations
- Hugely important driver of progress!
- Active research on improvements
- Hard to relate BLEU/NIST to some ordinary, easily comprehended assertion about MT quality

2003 Research Trend: From Word Translations to Phrase Translations



2004 Research Trend: Error-Based Training



Algorithms for Directly Reducing Translation Error

Some TIDES MT Research in 2004...

- Sentence alignment by top-down splitting
- Symmetrized word alignment
- Advanced Chinese segmentation
- Integration of external specialized translation components
- Partitioning parallel data for training on large data sets
- Advanced models of word-order change
- Language model adaptation
- Simplex method for tuning parameters
- Syntax-based translation models
- Improved MT evaluation metrics
- Weighted finite-state tools for decoding
- Extracting parallel sentences from comparable corpora

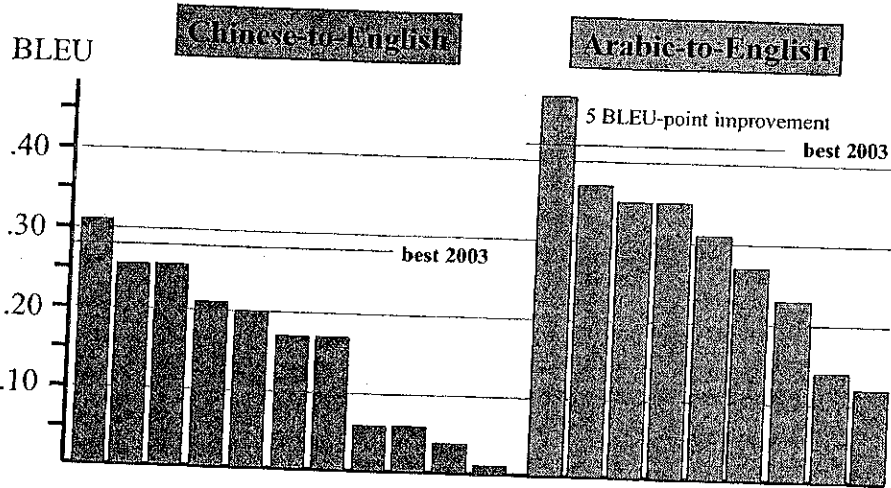
May 2004 Common Evaluations

- 3rd annual Chinese/English and Arabic/English evaluation
 - Administered by NIST
 - Scores immediately returned to participants
- 17 participants (last year: 10, previous: 7)

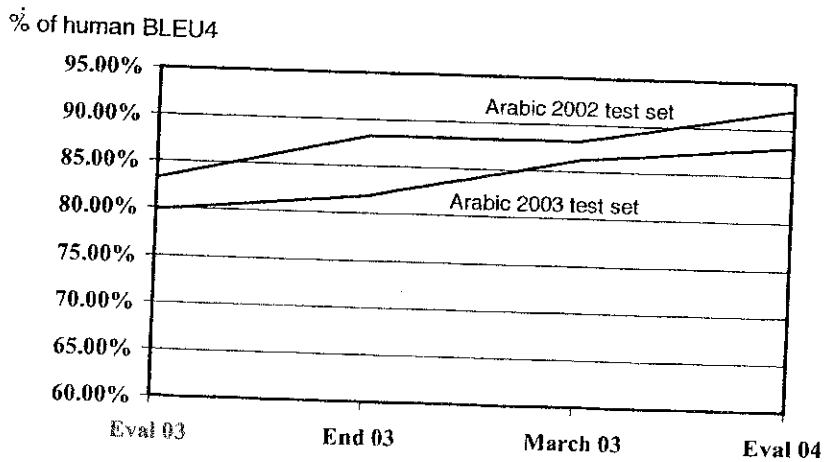
US Army Research Lab
ATR
Brigham Young University
Carnegie Mellon
First Capital Technologies
Harbin Institute of Technology
HKUST
IBM

ITC-irst
JHU/Maryland
Linear B
MIT
MTM Linguasoft
NTT
RWTH Aachen
StreamSage
USC/ISI

NIST 2004 MT Evaluations (News)



Percent of Human BLEU



TIDES Arabic MT Milestones

- 2002 result: 54% of human
- 2003 goal: 65% of human
- 2003 result: 80% of human
- 2004 goal: 85% of human
- 2004 result: 74% of human
- What happened?
- Good improvement in machine translation ☺
 - 5 BLEU points
- Even better improvement in human translation ☹
 - 11 BLEU points

Other Genres Added in 2004

	Arabic	Chinese
News	0.47	0.31
Editorials	0.35	0.29
Speeches (original text, not ASR)	0.42	0.35

BLEU scores

Last Year's Issues

- How can we build better machine translation algorithms?
- How can we get more data?

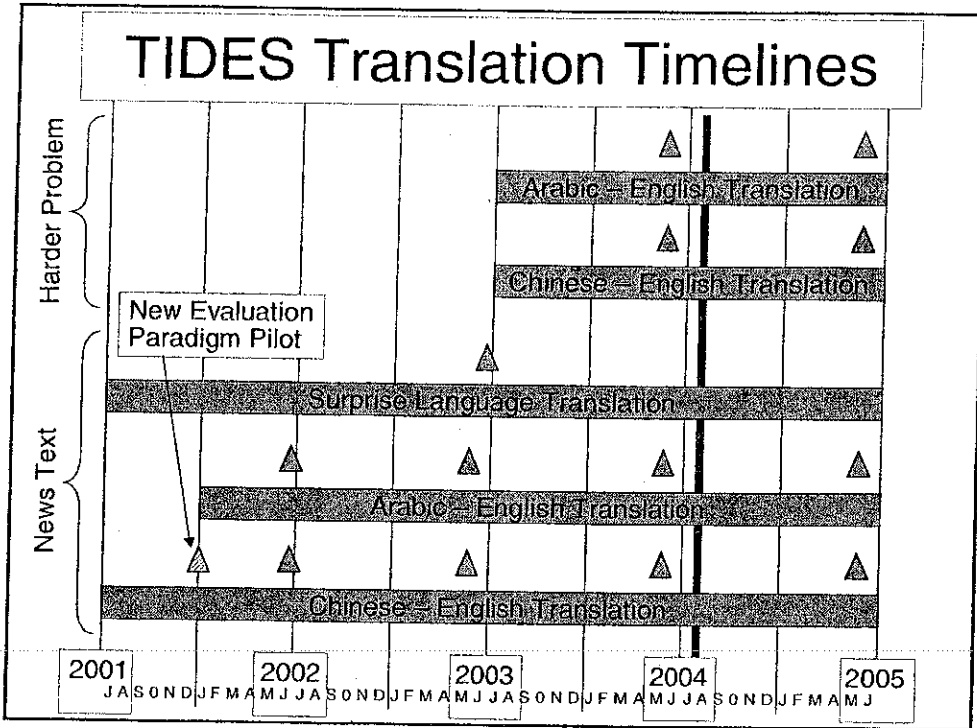
- Can we continue using BLEU to drive research progress?
 - YES
- Should we make the problem harder?
 - ADDED NEW GENRES, PERHAPS TOUGHER ONES NEEDED
- How to get more participants?
 - 7 → 10 → 17 → ...
- How can we characterize MT accuracy as a function of human performance?
 - Percent of Human BLEU

This Year's Issues

- How can we build better machine translation algorithms?
 - Lots of ideas, if NIST workshop is any indication
- How can we get more data?
 - "Data is outstripping computational resources & software..."
 - But that's good!

- Why did Percent of Human BLEU go down?
 - Why did Arabic human translators get so much more consistent – i.e., get higher human-human BLEU scores – between 2003 and 2004?
 - How much actual progress did MT make?
 - Was the 2004 test set easier?
 - Why are Arabic translators so much more consistent than Chinese translators? Is the same reason Arabic MT is better than Chinese MT?
- How can we characterize certain levels of MT accuracy as enabling (or not enabling) some application-level capability? (I.e., are we done yet?)
 - OK, that's a tough one.
 - Current studies underway at DLI/MIT-LL & others.

translation of news, local word reordering



Informal Poll on MT Research 2004

Why do you think Arabic/English MT performs better than Chinese/English MT?

- *RESPONDENT 1:* Chinese is very information-poor: things like prepositions, plural/singular have to be invented on the English side with only the help of the language model, which is quite hard.
- *RESPONDENT 2:* Arabic-English has a better match on word-meaning-word. Chinese requires more phrasal translation.
- *RESPONDENT 3:* Arabic systems yield a smaller search space during decoding → less opportunity to mess up.
- *RESPONDENT 4:* Word segmentation. In Chinese, UNK words are not really UNK words.
- *RESPONDENT 5:* $p(\text{English}|\text{Arabic})$ has less 'perplexity' than $p(\text{English}|\text{Chinese})$ → has to be like that because there are more Arabic words (because of richer morphology)
- *RESPONDENT 6:* Effort in tuning.
- *RESPONDENT 7:* My first reaction was that the Arabic source data must be more 'homogeneous' than the Chinese source data. My idea was that the models would be easier to train in this case, and there'd be more similarity between training and test. But this didn't pan out empirically.

What's the most effective role for native speakers of Chinese and Arabic in your MT system development?

- I did not have any input from native speakers. At this point I would like to have some help on ideas like morphology, number translation, simple normalizations, so in essence: preprocessing.
- Making sure that we convert the test data correctly. This year we had about 1 day effort by an Arabic native speaker. For Chinese it is difficult to say, as several students are Chinese. Work on named entities (Chinese) surely profits from language knowledge.
- Error analysis at all levels (alignment, decoding, etc).
- As consultant for doing error analysis.
- Writing rules for numbers/dates/bylines/...
- Identify model shortcoming by finding gross errors.
- Sanity checks for bitext training and input and output formatting.
- Tokenization issues.

What sub-topic doesn't have enough published papers?

- Everything. I'd especially like to read more about word-alignment and phrase extraction, with evaluation on machine translation performance.
- System details, i.e., the little things which are necessary to achieve the good results.
- Implementation details, especially for decoders.
- Discriminative training for machine translation.
- Syntax-based language models.
- Optimization of millions of parameters.
- How do humans translate?
- Alignment accuracy.
- The hidden side of the iceberg, i.e., engineering details and smart tricks to get good performing systems. This information are never published and is getting quite relevant in order to test (and possibly publish) new ideas on top of state-of-the-art technology.
- End-to-end system descriptions with detailed analysis of contributions of system components to overall performance.

What sub-topic has too many published papers?

- None.
- None.
- None in particular.
- Papers on evaluation metric X, and Y, and then there is Z, Z1, Z2 ;-)
- The sub-topic "random idea evaluated on non-standard and too small test corpus."
- MT evaluation.
- Yet another way to compute phrase probabilities.
- Yet another Bleu metric.
- Any NLP component (parser, tagger, etc.) not evaluated in the context of a complete translation task.

What is the single biggest technical problem holding back MT performance?

- A good way to deal with integrating syntactic knowledge.
- Again, no single biggest technical problem: MT performance will improve by making little improvements at all fronts. From the data we have we could generate much better translations, if only the models would provide a better ranking of the different translations.
- Computers are too slow; their RAMs are too small.
- Better translation and language models.
- Better search algorithms.
- Inadequate hardware resources and software infrastructure at MT research labs cannot handle huge amounts of data available nowadays.
- more
- To compute good alignments from very large and noisy parallel corpora.
- Absence of techniques that can benefit from large bitexts. Right now the performance of systems seems to taper off rather quickly; the gains from adding more data in training are small.

What year will MT reach average human BLEU4 performance?

- I don't care. I don't think human BLEUx scores are useful.
- 2010 for easier languages, 2015 for more difficult languages.
- 2007.
- On Arabic 2002 test data: it will happen by next year.
- In general: unclear, because answer depends on variance of those four human translators which seems to vary very much from year to year.
- 2008 on a broad multi-genre test set.
- 2001 could have been the right year ... maybe in five years, but I fear that getting close to human BLEU scores will just show the limitations of BLEU.
- MT is already better than the average human at translating.

What do you think is the single most useful data resource currently available?

- All data is good data. Parallel corpora are more useful than dictionaries.
- Don't think there is any single most useful data resource, as it depends on what system I want to build. So, rather generally: bilingual corpora with good translation quality.
- UN corpus.
- In general for doing MT research: the TIDES MT training/test data for Chinese/Arabic.
- UN parallel corpora.
- Parallel corpora of good quality, i.e., with high quality translations.
- Chinese: LDC C-E dictionary.
- FBIS corpus.

What would you like to see happen on future data development for MT?

- More variety in corpora to deal with domain adaptation.
- More data like the UN corpora, big and of good quality. For speech translation: bilingual corpora consisting of recordings of original speech and interpreter, perhaps also transcription or ASR output for both languages.
- Word-level alignments (bilingual and multilingual).
- Parallel data from various domains.
- A large scale effort to get >1 billion words for one language pair.
- Many language pairs with >100 million words.
- Acquisition of more parallel corpora from different domains/genre
- Production of parallel corpora of better quality and reference alignments to evaluate training algorithms. To distribute alignments of training data computed by the best performing systems.
- In-domain bitexts.
- NIST should develop significance tests for BLEU scores, and, if necessary, the test sets should be increased in size so that we can determine if new methods are significant at the current state of the art.