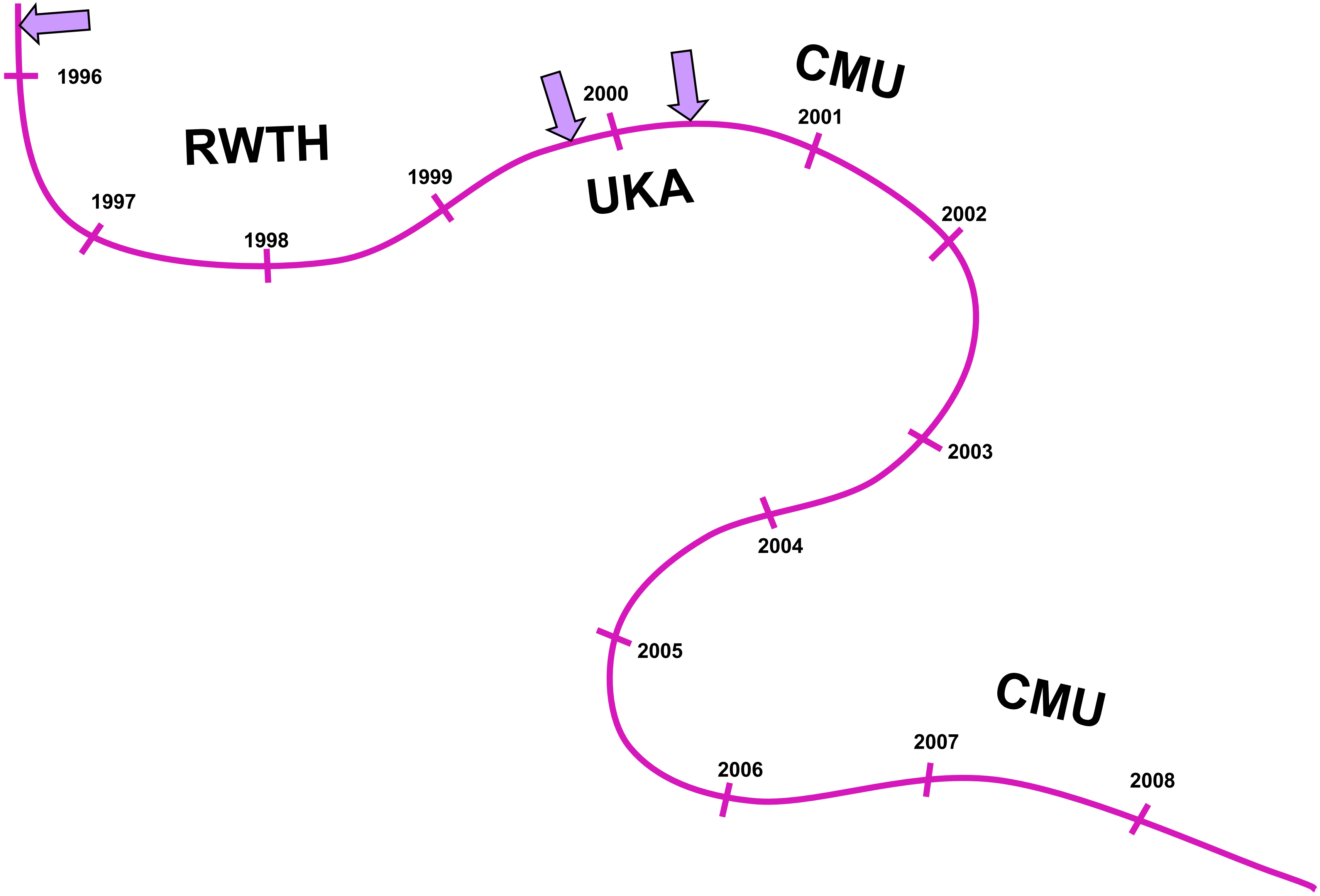
Speech Translation
From Domain-Limited to Domain Unlimited Translation Tasks

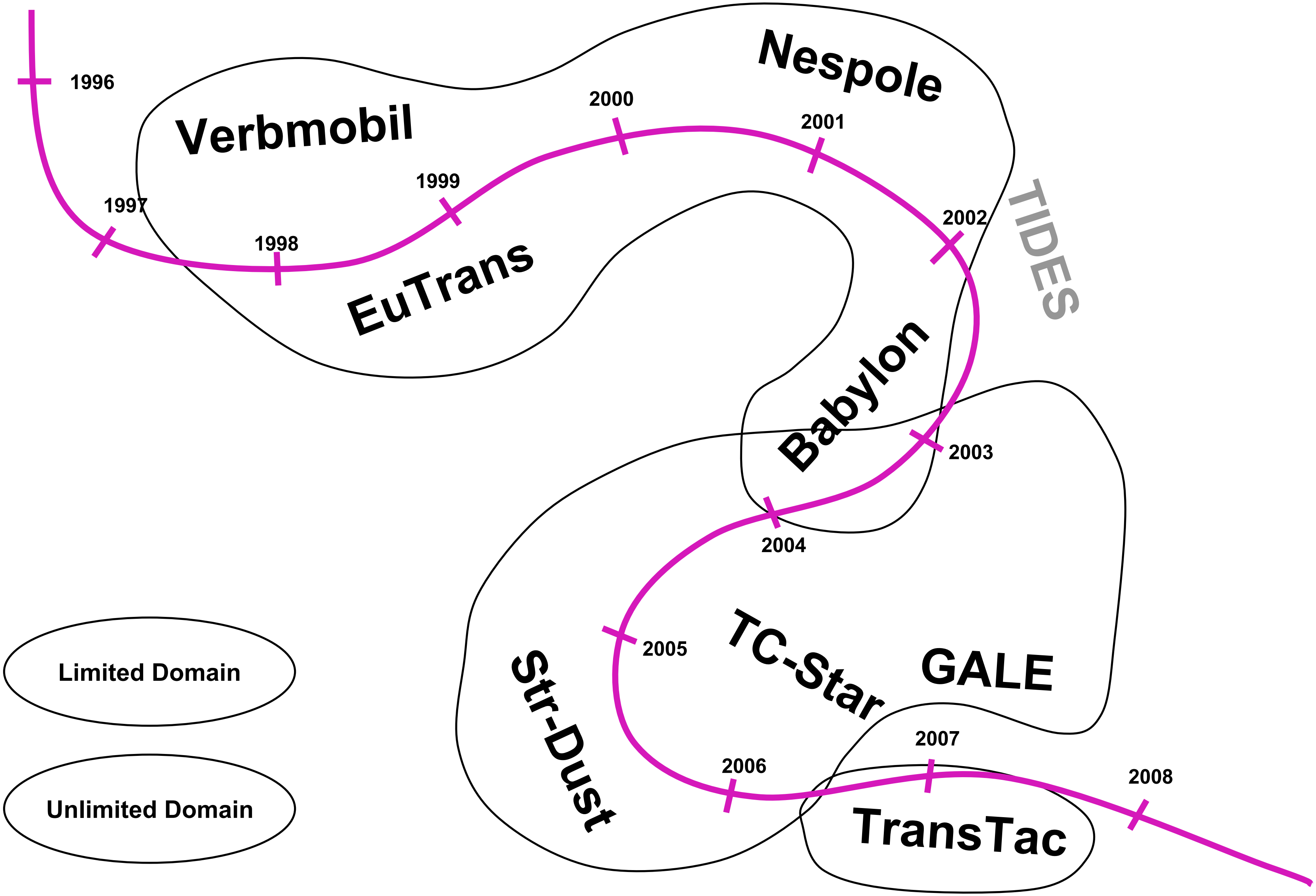
Stephan Vogel

InterACT
Language Technologies Institute
Carnegie Mellon

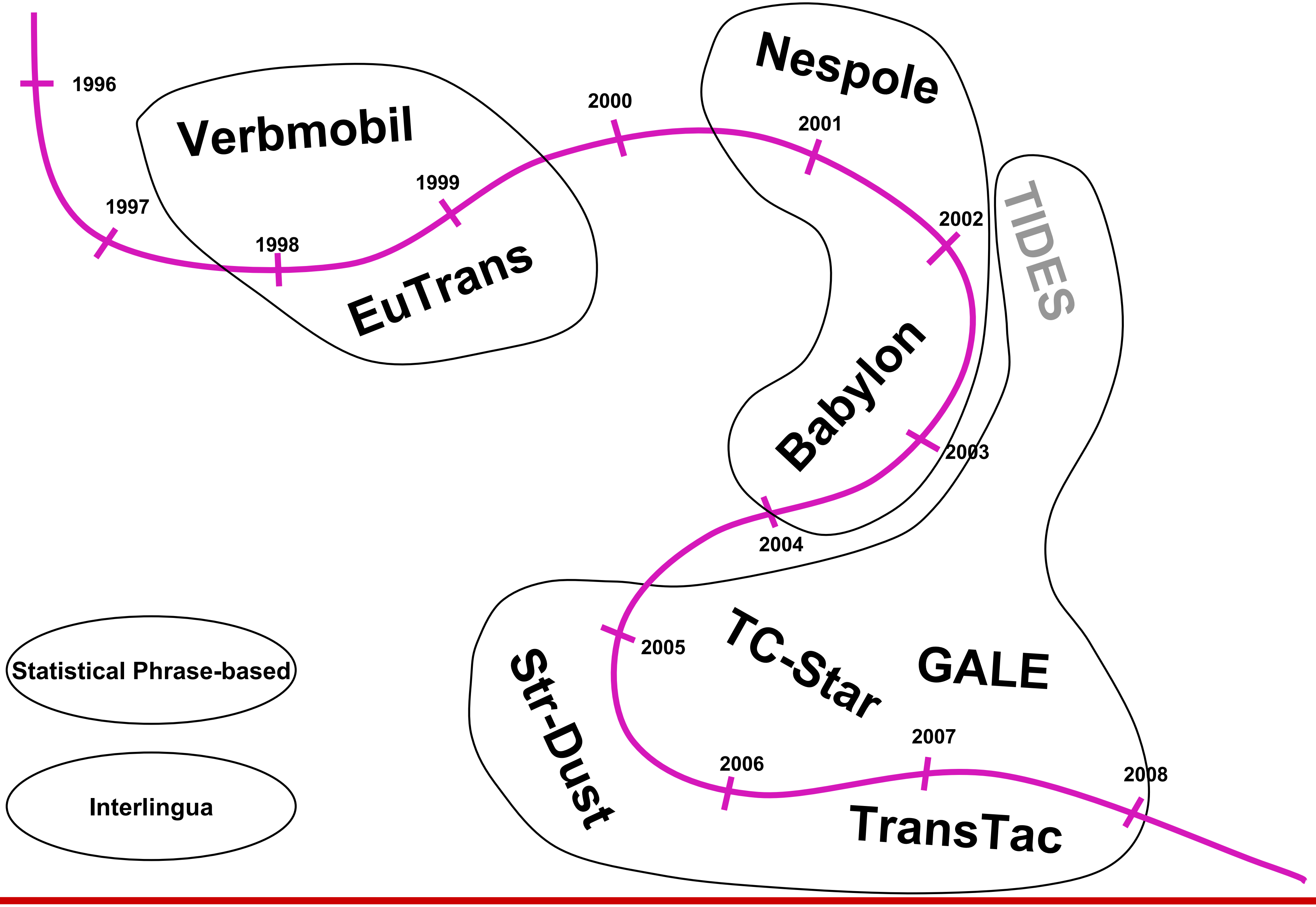
A (Personal) Speech Translation Journey



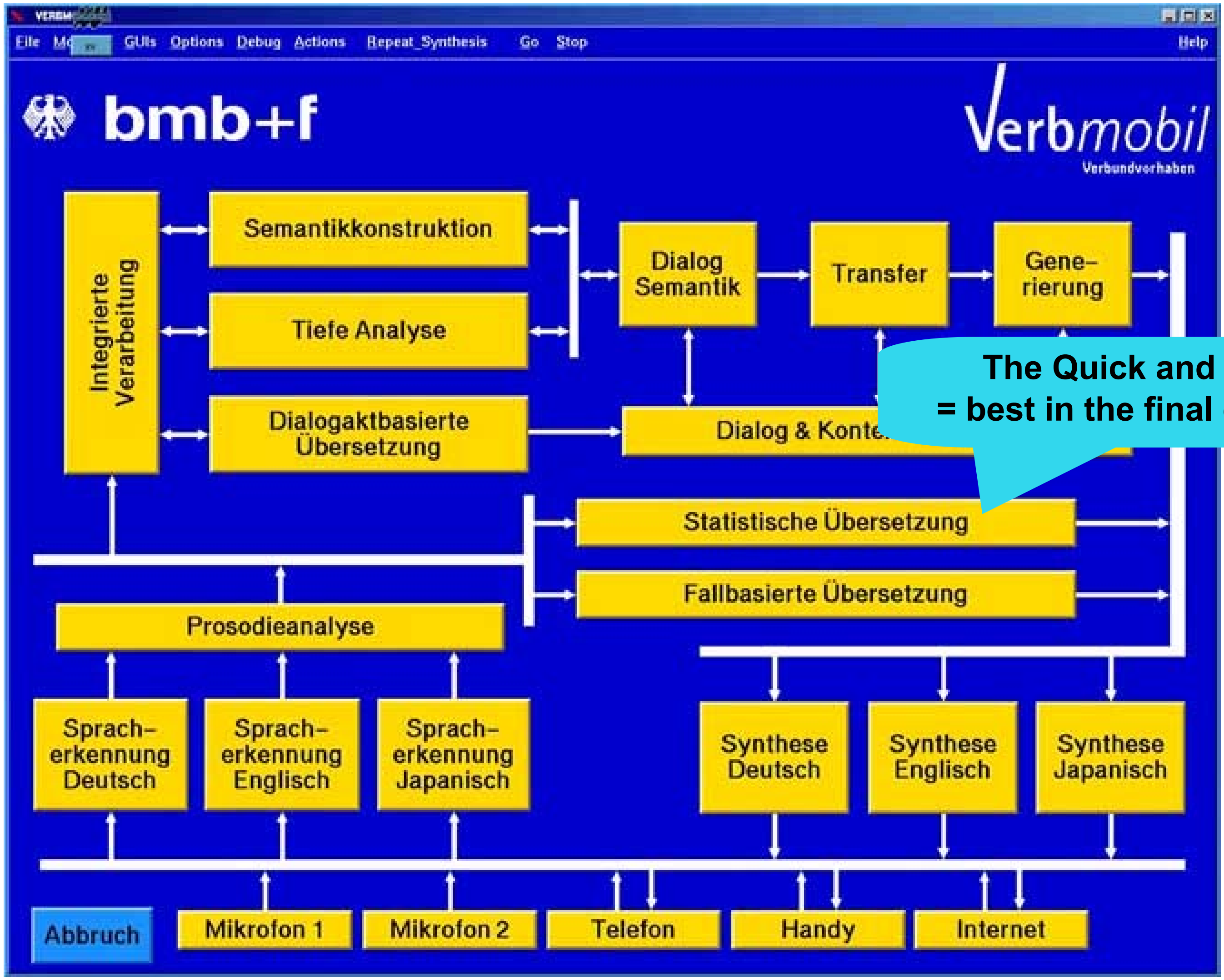
A (Personal) Speech Translation Journey



A (Personal) Speech Translation Journey



The Verbmobil Project



The Quick and Dirty ☺
= best in the final evaluation

Verbmobil Highlights

- Multiple recognizers (languages and systems)
- Multiple translation engines
  - Transfer, with deep syntactic/semantic analysis
  - Dialog-Act-based
  - Example based
  - Statistical
- System combination for translation engines
  - Did not work
- Prosodic annotation (probabilistic, on lattices)
  - Segment boundaries on various levels
  - Question
  - Accent
  - Actually used by translation modules
- Disfluency detection and removal (on lattices)

Nespole!

- Project
  - C-Star partners
  - Interlingua-based translation (English, German, Italian)
  - Travel domain
- Outside of project: test, if SMT is possible on very small corpora

Train: 3182 parallel speech dialog units (sentences)

Language	English	German
Tokens	15572	14992
Vocabulary	1032	1338
Singletons	404	620

Test: 70 Parallel SDUs

	German	Reference A	Reference B
Tokens	437	610	607
Vocabulary	183 (45 oov)	165	160

Evaluation

- Human Scoring
  - Good, Okay, Bad (c.f. Nespole evaluation)
  - Collapsed into a „human score“ on [0,1] (good = 1.0, okay = 0.5)
- Bleu Score
  - Average of N-gram precisions from (1..N), typically N=3 or 4
  - Penalty for short translations to substitute for recall measure
  - Numbers ranging 0.0 ... 1.0, higher is better

		Good	Okay	Bad	Score	Bleu
Text	IF	77	104	227	0,32	0,068
	SMT	127	80	205	0,40	0,333
Speech	IF	64	101	243	0,28	0,059
	SMT	95	83	227	0,34	0,262

- Results
  - SMT works (as good) with very small training data
  - Both systems pretty useless (OOV!!!)

Translation System for Travel Domain

Central question:

How to build a translation system which can be used when you are “On the Run”

Translation System for Travel Domain

- (Phrase-based) SMT is soooo greedy
- Memory
  - Big phrase tables
  - Large N-gram language model
- Computation
  - Decoder needs to generate and evaluate many alternative translations
- Typical machine for SMT systems
  - 2GHz CPU, 4++GB memory
    Linux machine

Central question:

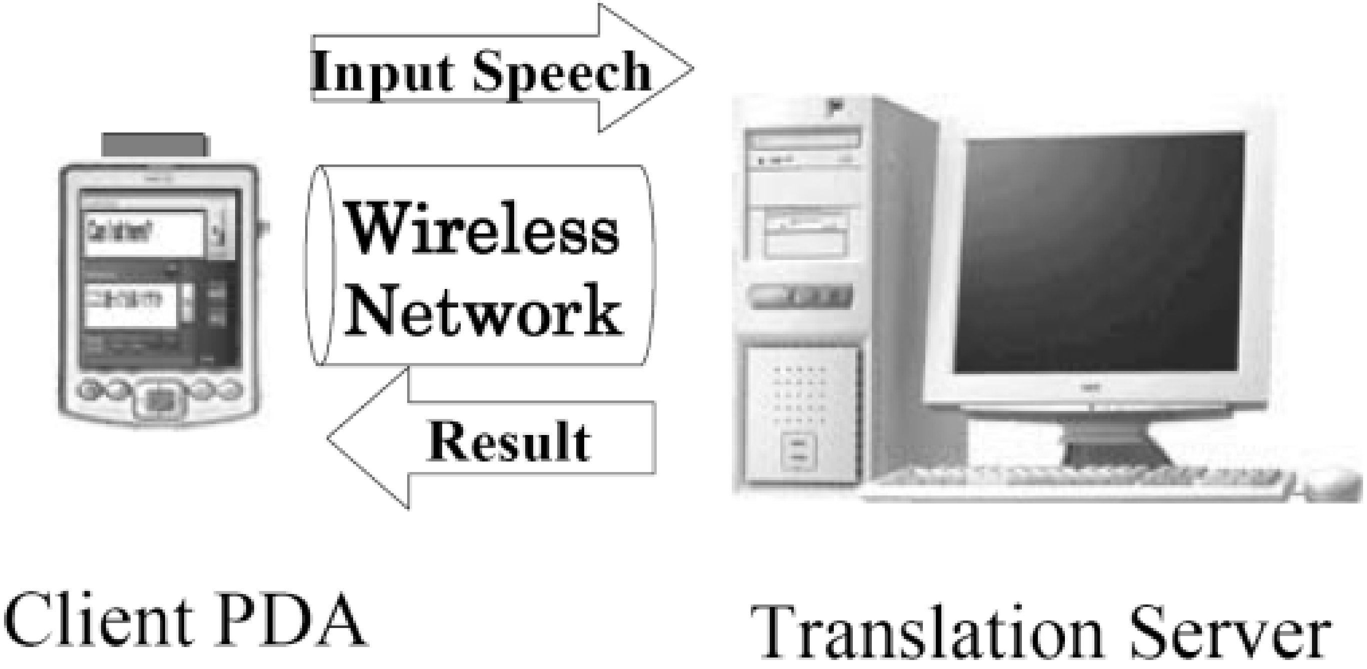
How to build a translation system which can be used when you are “On the Run”

Option I

- Travel with powerful computers

Option II

- Communicate with a powerful server via wireless*

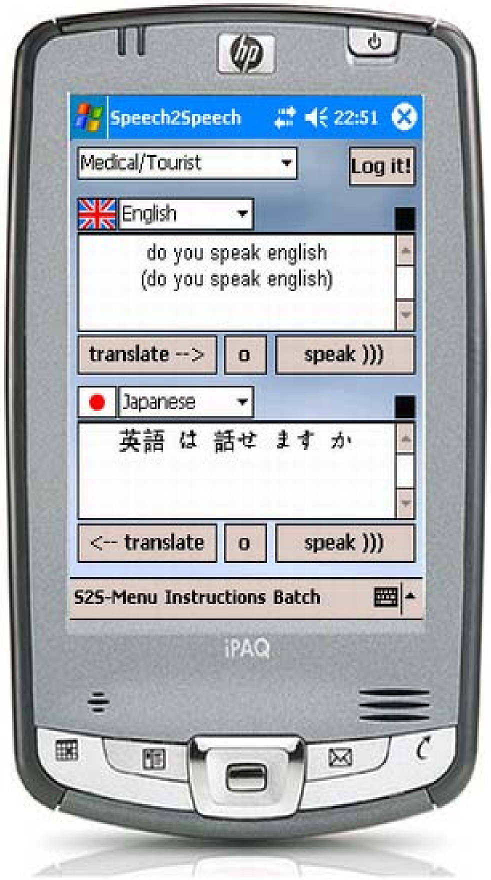


- Problem: Wireless networks are not always available

* Yamabana et al. 2003. *A speech translation system with mobile wireless clients*.

Option III: PanDoRA

- Smart engineering of the SMT system so that it can run on hand-held devices
- All components on device
  - No dependency on connectivity



Speech2Speech 22:51

Medical/Tourist

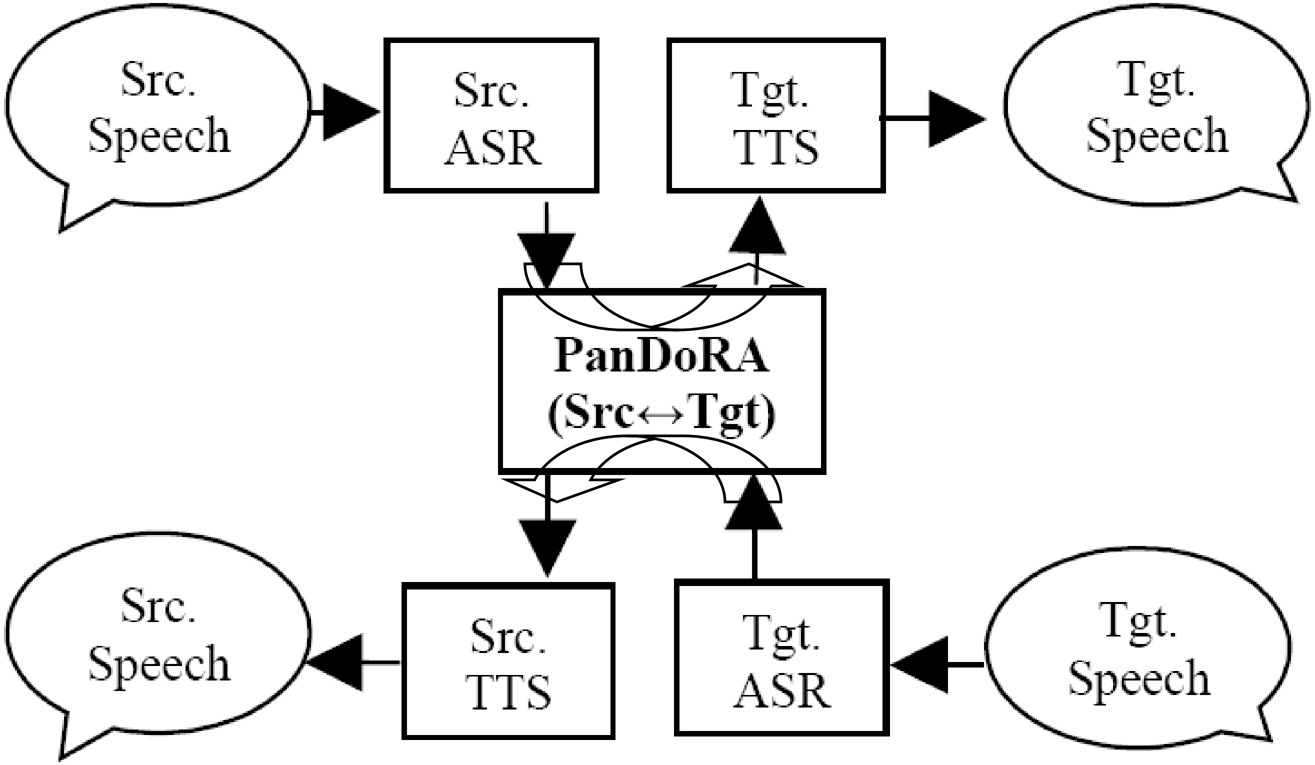
English

do you speak english
(do you speak english)

Japanese

英語 は 話せ ます か

PanDoRA: System Architecture



Challenges

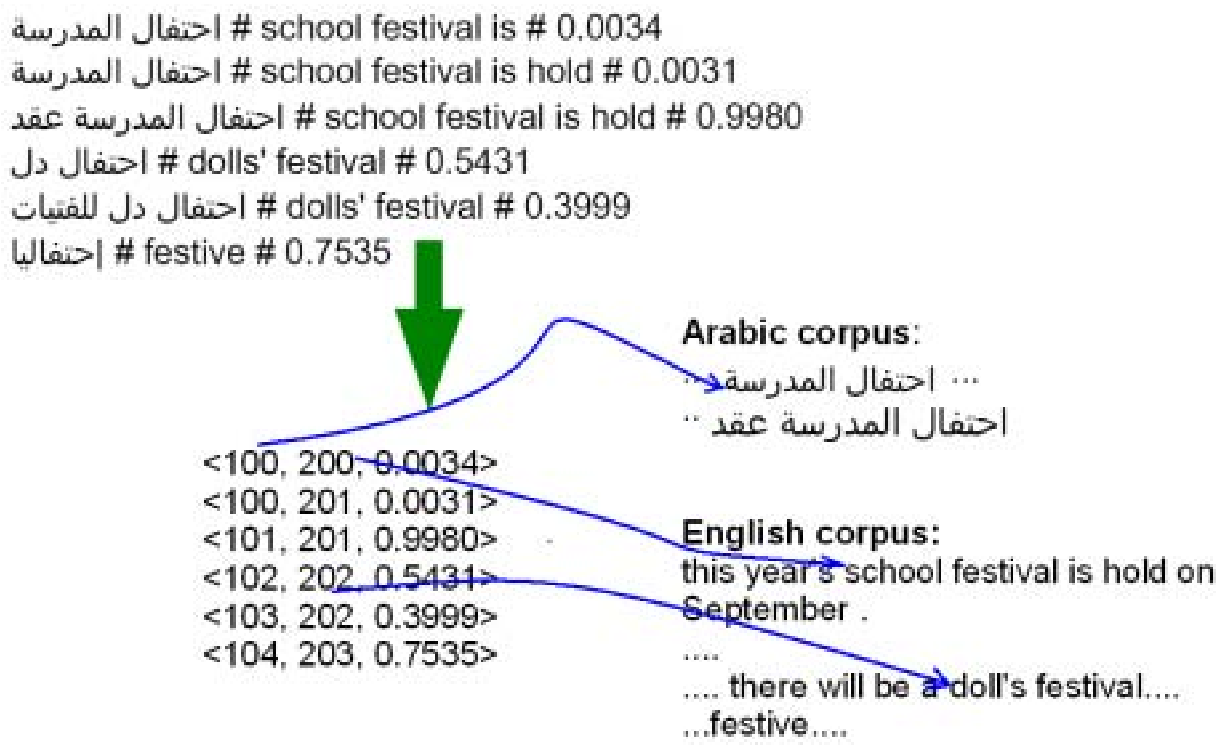
- Memory
  - SMT systems require several GB RAM on PCs
  - Hand-held devices have very small dynamic memory
  - 64MB available on iPaq 2750; shared with ASR and TTS
- CPU
  - Speech translator requires real-time translations
  - CPUs on hand-held devices are usually slow (e.g. 600M Hz) and
  - There are no coprocessors for floating-point arithmetic

Solutions

- Compact data structures
- Integerized computation
- Efficient decoding
- Minimum on-device computation

Compact Data Structure

- Each phrase is represented by its <location,length> in the corpus rather than a string of words
- A translation pair is represented by <src phrase index, tgt phrase index, score>



احتفال المدرسة # school festival is # 0.0034
احتفال المدرسة # school festival is hold # 0.0031
احتفال المدرسة عقد # school festival is hold # 0.9980
احتفال دل # dolls' festival # 0.5431
احتفال دل للفتيات # dolls' festival # 0.3999
احتفاليا| # festive # 0.7535

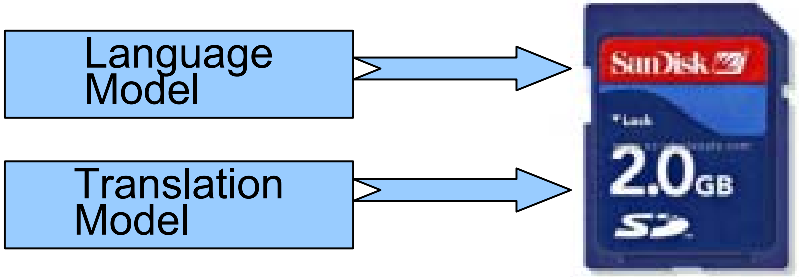
<100, 200, 0.0034>
<100, 201, 0.0031>
<101, 201, 0.9980>
<102, 202, 0.5431>
<103, 202, 0.3999>
<104, 203, 0.7535>

Arabic corpus:
... احتفال المدرسة ..
.. احتفال المدرسة عقد

English corpus:
this year's school festival is hold on September .
....
.... there will be a doll's festival....
...festive....

Compact Data Structure

- Data records now have fixed size in bytes, independent of the length of a phrase.
- Sorted list of all model data is stored on external memory card
  - Much larger capacity than SDRAM. e.g. 2GB SD card
  - Saves SDRAM for decoding
- Search for matched phrases by binary search in the sorted list
- Access the information by 'seek'ing the record from a file by record id.
  - e.g. getPhrase(id); getTranslationOfPhrase(id)...



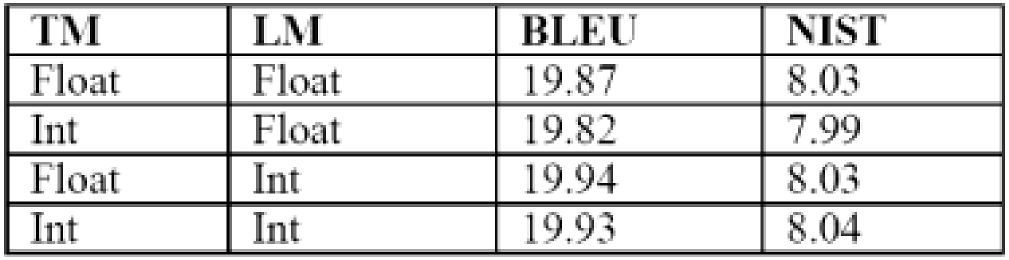
Language Model

Translation Model

- With compact data structure a 5.6 million entry phrase table requires 65MB on disk
- Specification for current implementation
  - 64K Vocabulary for each language
  - Up to 256 million unique src/tgt phrases allowed
  - Up to 4 billion phrase pairs allowed

Integerized Computing

- SMT systems are “translating by numbers”
- Hand-held devices usually do not have numerical co processors
- Slow with floating points calculations
- PanDoRA's solution: integerized computing
  - Convert all probabilities to cost = -log(prob)
  - Quantize the cost to integer bins between [0, 4095]
  - Probabilities put into the same bin are considered equal during decoding.



TM	LM	BLEU	NIST
Float	Float	19.87	8.03
Int	Float	19.82	7.99
Float	Int	19.94	8.03
Int	Int	19.93	8.04

Minimum On-device Computation

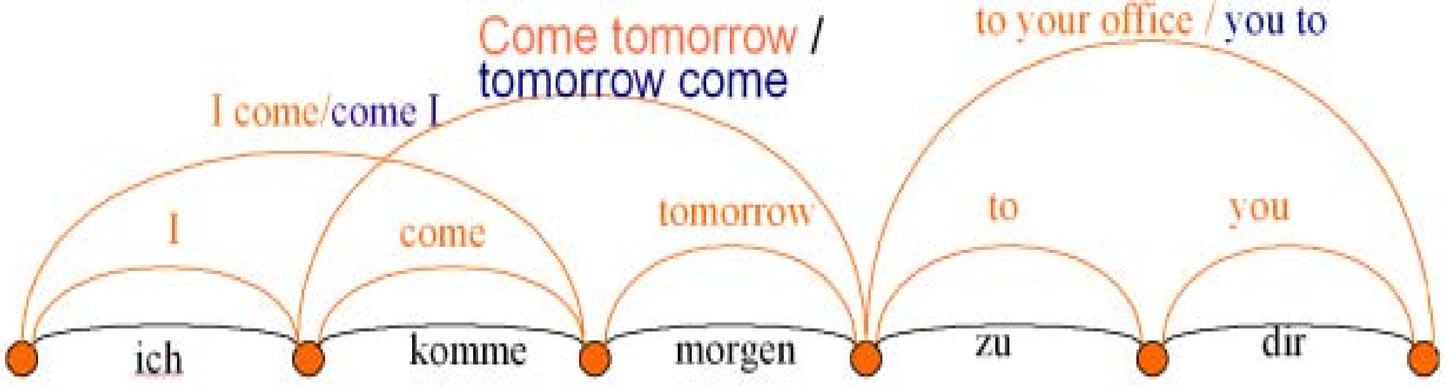
- Log-linear model in SMT
  - Multiple model features

$$P(e_1^I \mid f_1^J) = \frac{\exp(\sum_{m=1}^{M} \lambda_m \phi_m (e_1^I, f_1^J))}{Z}$$

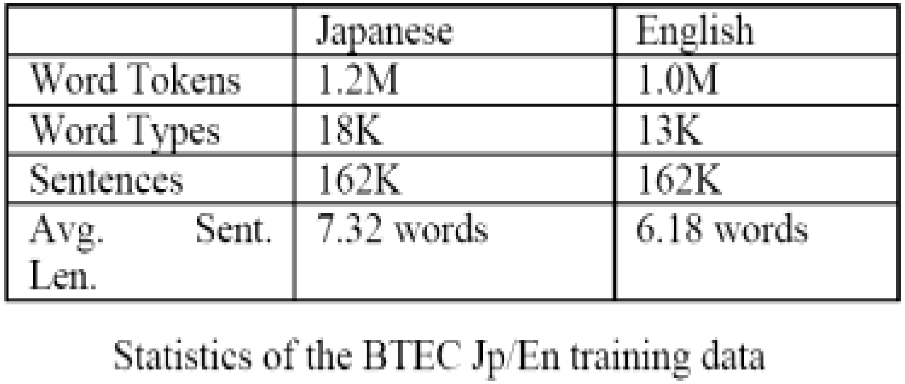
- Off-line model optimization
  - Minimum Error (MER) training off-line
  - Combine weighted translation model scores to a single score

Decoding

- Inverted Transduction Grammar (ITG) style decoding [bottom-up]
  - Translation as parsing
  - Inverted Transduction Grammar (ITG) (Wu, 1997)
    - Straight transduction: X-><f1f2 | e1e2>
    - Invert transduction: X-><f1f2 | e2e1>
  - Combine adjacent partial hypotheses either straight or inverted to create new hypothesis covering the combined source range.
  - Allows for long distance reordering. Effective for language pairs with dramatically different word ordering. E.g. Japanese/English
  - Language model adjusted on the boundary words.

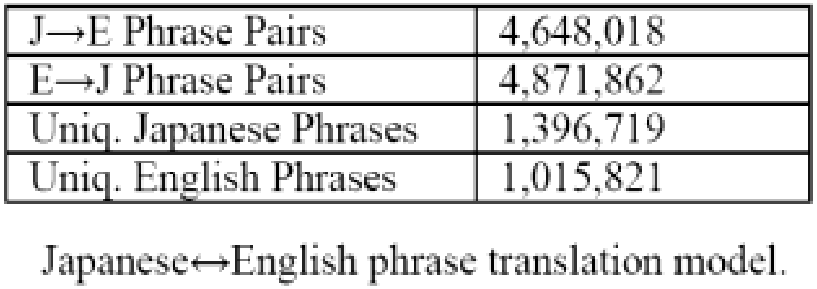


Experiments on Japanese/English



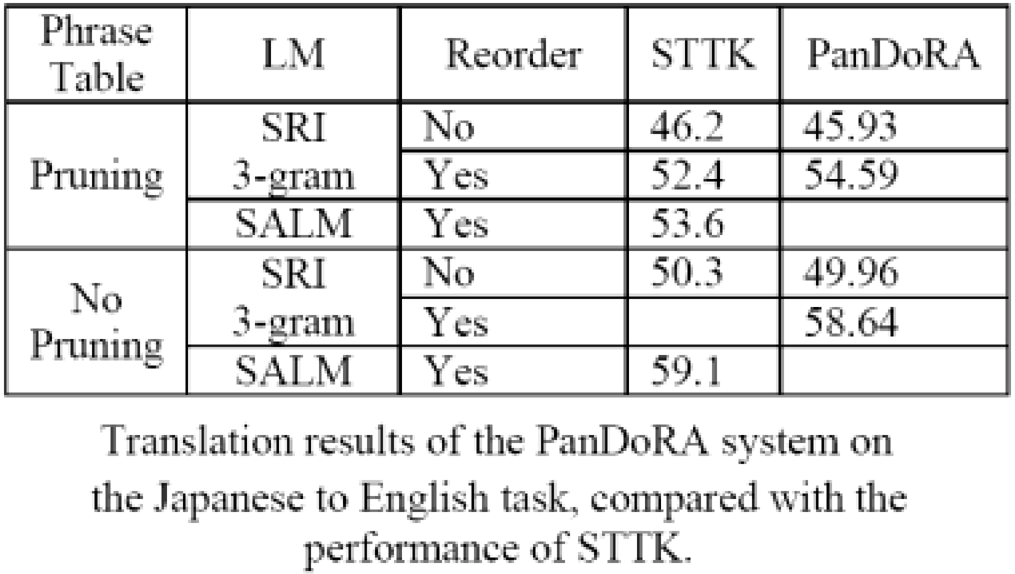
	Japanese	English
Word Tokens	1.2M	1.0M
Word Types	18K	13K
Sentences	162K	162K
Avg. Sent. Len.	7.32 words	6.18 words

Statistics of the BTEC Jp/En training data



J→E Phrase Pairs	4,648,018
E→J Phrase Pairs	4,871,862
Uniq. Japanese Phrases	1,396,719
Uniq. English Phrases	1,015,821

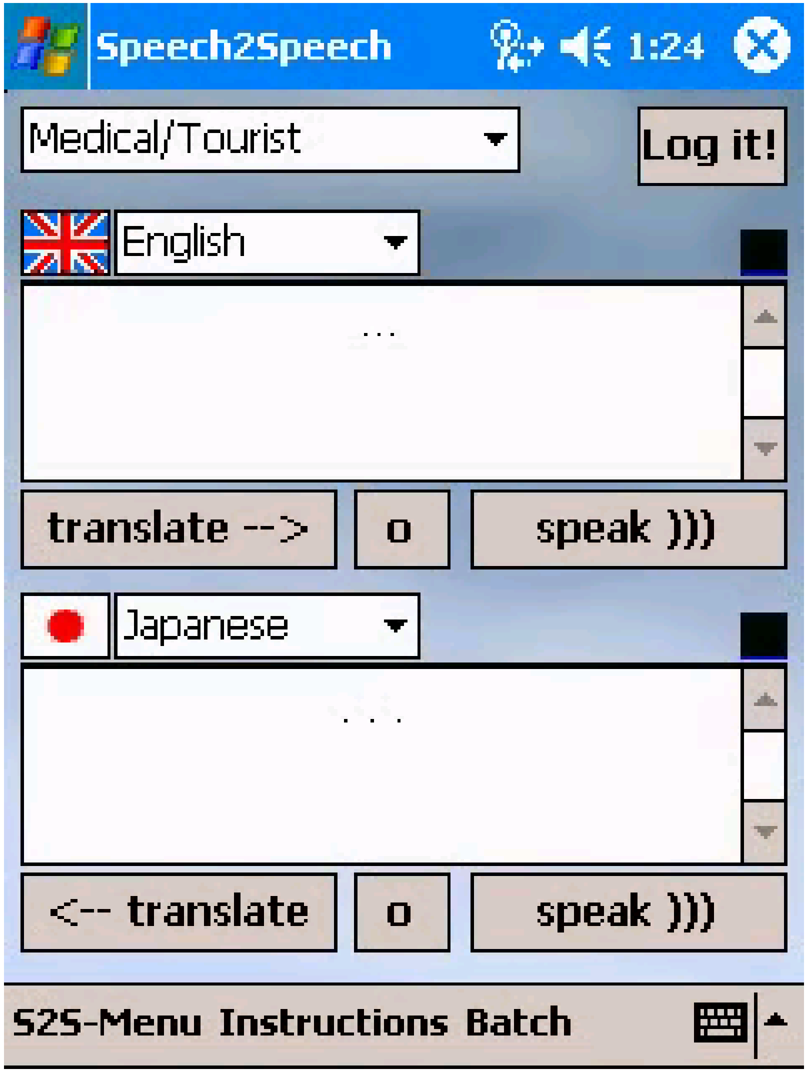
Japanese↔English phrase translation model.



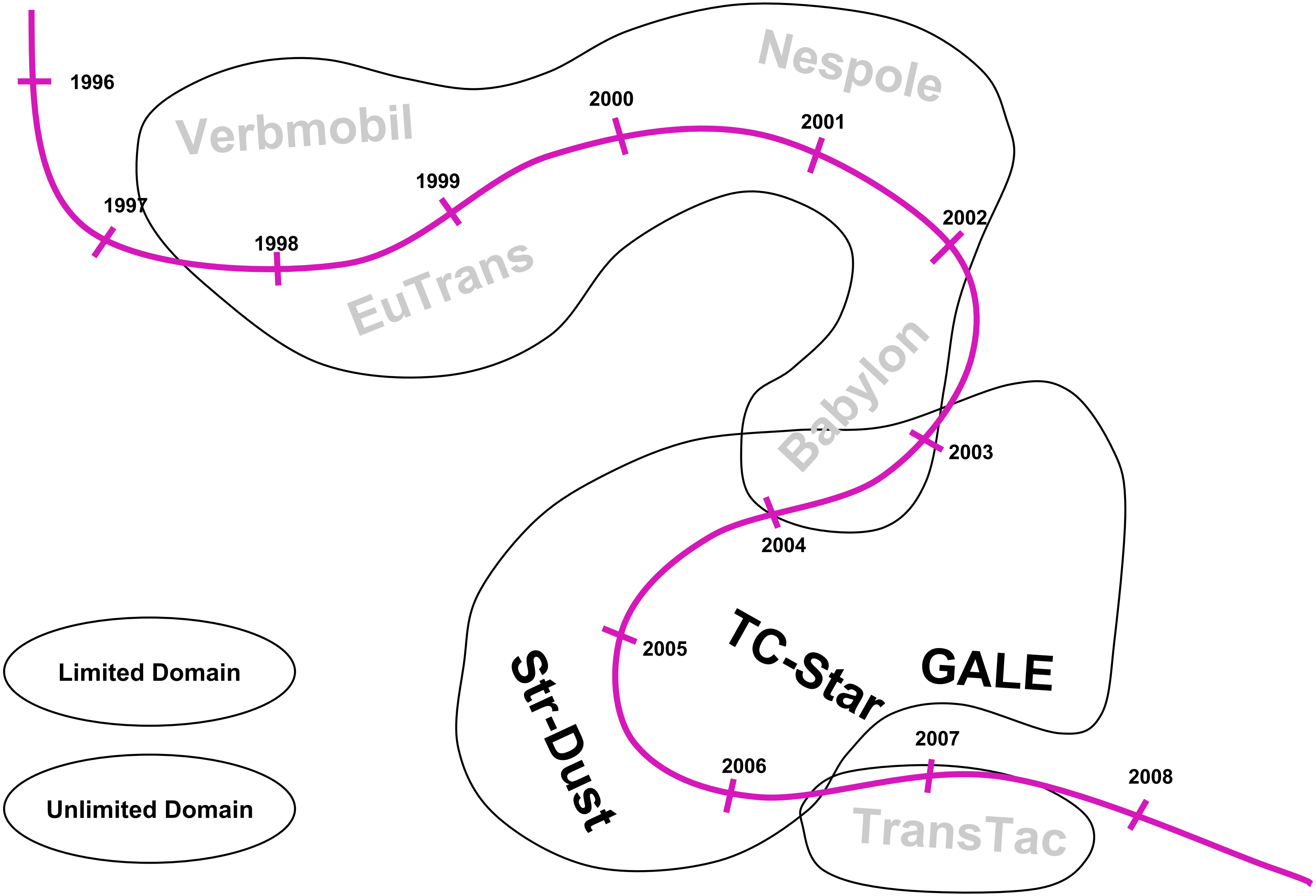
Phrase Table	LM	Reorder	STTK	PanDoRA
Pruning	SRI 3-gram	No	46.2	45.93
		Yes	52.4	54.59
	SALM	Yes	53.6	
No Pruning	SRI 3-gram	No	50.3	49.96
		Yes		58.64
	SALM	Yes	59.1	

Translation results of the PanDoRA system on the Japanese to English task, compared with the performance of STTK.

Travel Domain Translation Demo



From Limited towards Unlimited



TC-Star

Technology and Corpora for Speech to Speech Translation

- Financed by European Commission within the Sixth Program
- Research in all core technologies for Speech-to-Speech Translation

- Long-term research goals of the project:
  - Effective Spoken Language Translation of unrestricted conversational speech on large domains of discourse.
  - Speech recognition able to adapt to and perform reliably under varying speaking styles, recording conditions, and for different user communities.
  - Effective integration of speech recognition and translation into a unique statistically sound framework. (Lattice translations as explicit goal)
  - General expressive speech synthesis imitating the human voice. (Development of new models for prosody, emotions and expressive speech.)

TC-Star Evaluations

- EPPS: European Parliamentary Plenary Sessions)
  - EU >20 official languages
  - Speeches given in native language
  - Simultaneous interpretation (verbatim transcript)
  - Final text edition for publication (often significantly edited)

- TC-Star Evaluations
  - Show-case for potential translation (interpretation) service in the European Parliament
  - Spanish-to-English and English-to-Spanish
    - Verbatim transcript
    - ASR output (1-best, lattice)
    - Final text edition
  - Also Chinese-to-English (NIST MT eval clone?)

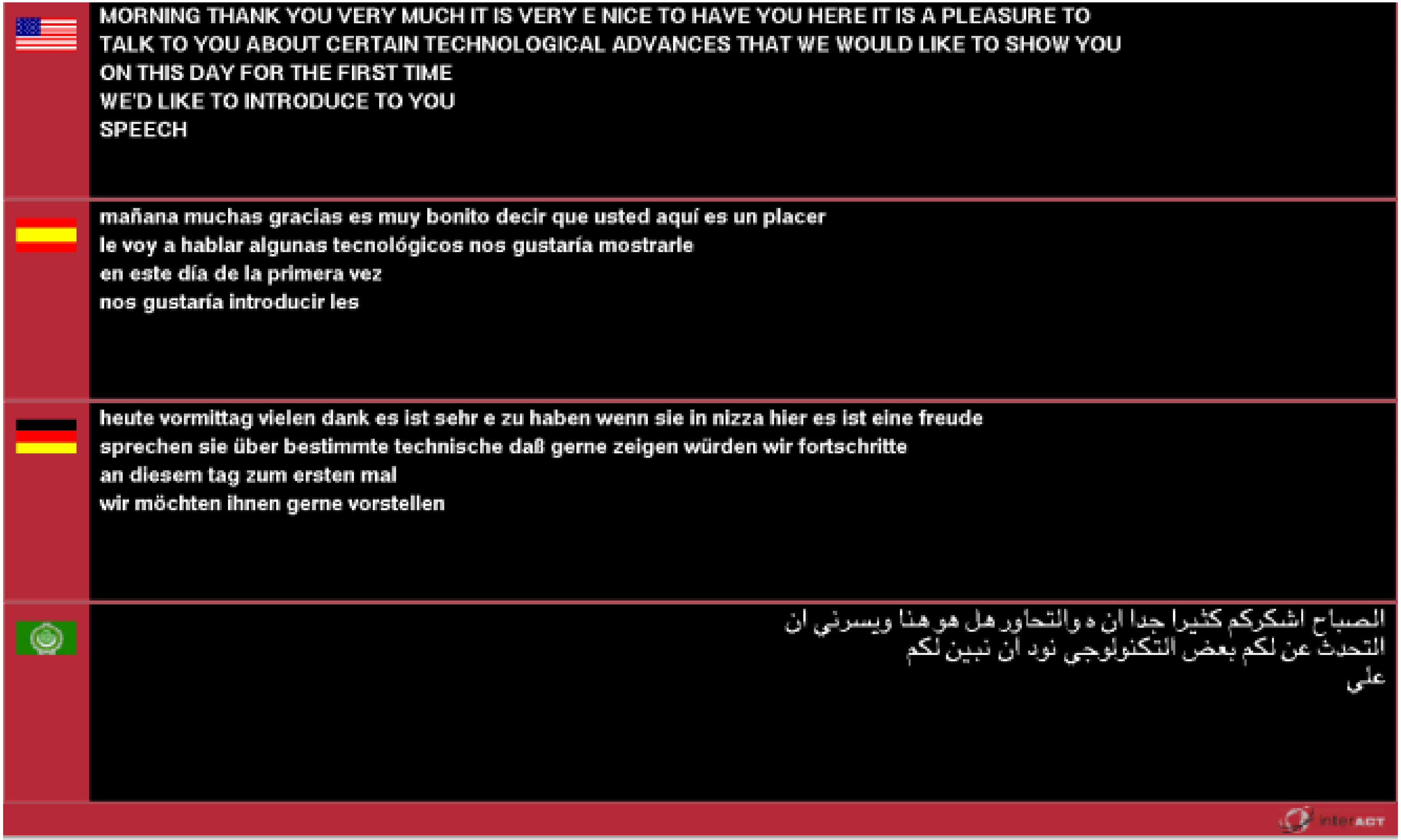
TC-Star Demo

Lecture Translator

- Speech recognition
  - Large vocabulary
  - Continuous speech
  - Fair amount of disfluencies
- Translation
  - English -> Spanish
  - English -> German
    - Trained on EPPS (European parliament) data
  - English -> Arabic
    - Trained on UN corpus
  - Additional LM data for adaptation to domain
- Application/Demo
  - Translating lectures, e.g on speech and translation technology

Lecture Translation Demo

Lecture Translation Demo



MORNING THANK YOU VERY MUCH IT IS VERY E NICE TO HAVE YOU HERE IT IS A PLEASURE TO
TALK TO YOU ABOUT CERTAIN TECHNOLOGICAL ADVANCES THAT WE WOULD LIKE TO SHOW YOU
ON THIS DAY FOR THE FIRST TIME
WE'D LIKE TO INTRODUCE TO YOU
SPEECH

mañana muchas gracias es muy bonito decir que usted aquí es un placer
le voy a hablar algunas tecnológicos nos gustaría mostrarle
en este día de la primera vez
nos gustaría introducir les

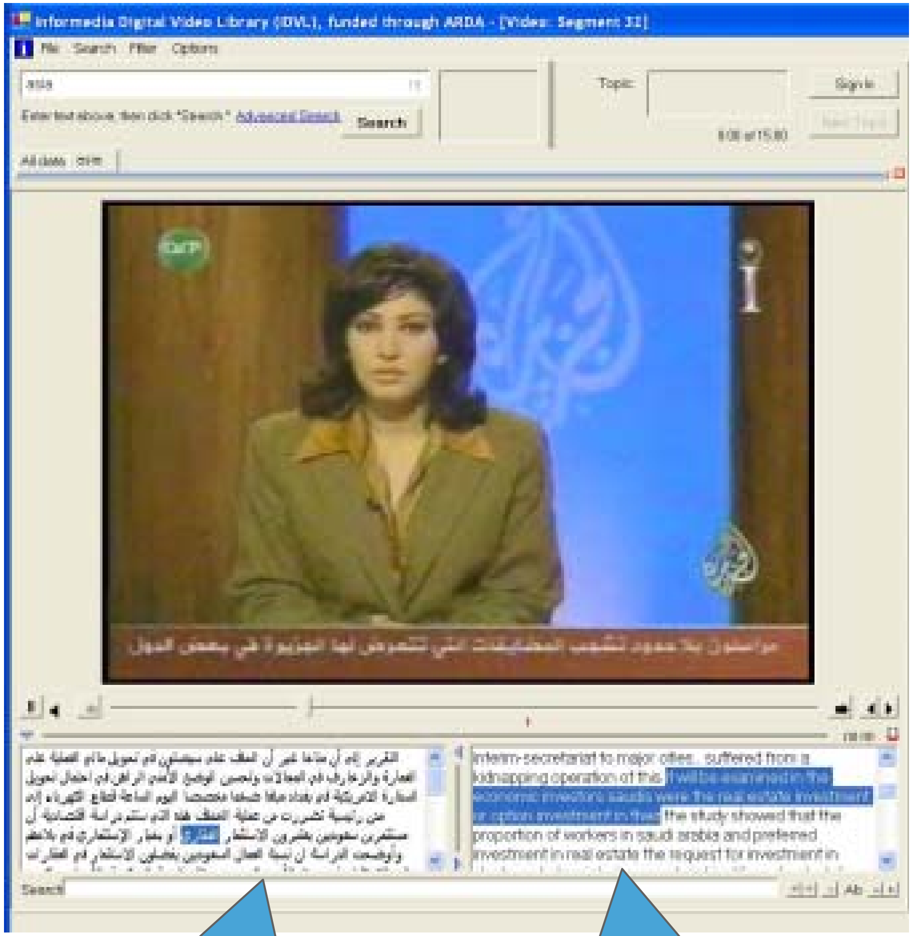
heute vormittag vielen dank es ist sehr e zu haben wenn sie in nizza hier es ist eine freude
sprechen sie über bestimmte technische daß gerne zeigen würden wir fortschritte
an diesem tag zum ersten mal
wir möchten ihnen gerne vorstellen

الصباح اشكركم كثيرا جدا ان ه والتحاور هل هو هنا ويسرني ان
التحدث عن لكم بعض التكنولوجي نود ان تبين لكم
على

Translation Server for Informedia

- Informedia system captures videos
- Sound track is send to ASR server
- 1-best result send to translation server
- Results archived in Informedia database
- Highlighting in text synchronized with video

..¥Demos¥InterACT-BNT¥InterACT-BNT.exe



Informedia Digital Video Library (IDVL), funded through ARDA - [Video: Segment 32]

File Search Filter Options
asia
Enter text above, then click "Search." Advanced Search
Search
Topic:
Sign In
All dates: 5hrs
Search

[illegible]

interim secretariat to major cities . suffered from a kidnapping operation of this . it will be examined in the economic investors saudis were the real estate investment or option investment in this the study showed that the proportion of workers in saudi arabia and preferred investment in real estate the request for investment in

Recognized Arabic

Translation into English

Some Research Topics

- Tight coupling between recognition and translation
  - Lattice translation
  - Using ASR features (LM, acoustic scores, confidences) in MT decoder
  - End-to-End optimization
- Segmentation and restoring punctuation
- Disfluency detection and removal

Lattice Translation: Motivation

- Lattice translation is used for tighter coupling between speech recognition and translation
  - Use alternative word sequences
  - Lattice is more efficient than n-best list
  - Word lattice has lower word error rate than 1-best hypothesis -> perhaps we find a better path in translation
  - Perhaps a different path gives better translation even when more recognition errors
  - Decouple processing steps, yet use relevant information for end-to-end optimization

- Lattices can be used to encode alternatives arising from other knowledge sources
  - Disfluency annotation: add edges to lattice to skip over disfluencies
  - Add synonyms and paraphrases
  - Add morphological variants, esp. for unknown words
  - Allow different word segmentation (e.g. For Chinese)
  - Partial word reordering

Requirement: Translating Lattices

- Connection between speech recognizer and translation system can be
  - 1-best hypothesis
  - N-best hypotheses list
  - Word lattice
  - Confusion lattice

- Why lattice
  - Compact representation of many hypotheses
  - Word lattice has lower word error rate than 1-best hypothesis
    -> perhaps we find a better path in translation
  - Perhaps a different path gives better translation even if it has more word errors
  - Can use results from consolidation modules without making a hard decision: consolidation module adds paths to the lattice

Requirement: Using ASR Features

- Problem when translating lattices
  - Perhaps bad path is chosen because it is “easier” to translate
  - Shorter paths are more likely to be selected
- Solution: Use Recognizer Features
  - Acoustic scores: local to source word
  - Language model scores: non-local
    - Expansion of the word lattice for different histories
    - Calculation on the fly, i.e. in the SMT decoder
    - Affected by word reordering
  - Confidence scores
- Other features might be interesting
  - Segmentation: probabilities for boundary after each word
  - Consolidation modules: probabilities for new edges
- Goal: overall optimization of end-to-end system

Requirement: Word Reordering

- Long distance reordering
  - Allow first word to be translated last, last word to be translated first
  - All permutations is too expensive
  - Typically only a small number of reorderings in a sentence
- Types of reordering
  - Jump ahead a few words, fill gap one or two steps later (only local word reordering)
  - Leave small number of gaps, fill at any time (global reordering possible)
  - Binary Tree with swapping sub-trees: not all reorderings orderings possible, but pretty much all which appear in natural languages (global reordering possible)

Different Reordering Strategies

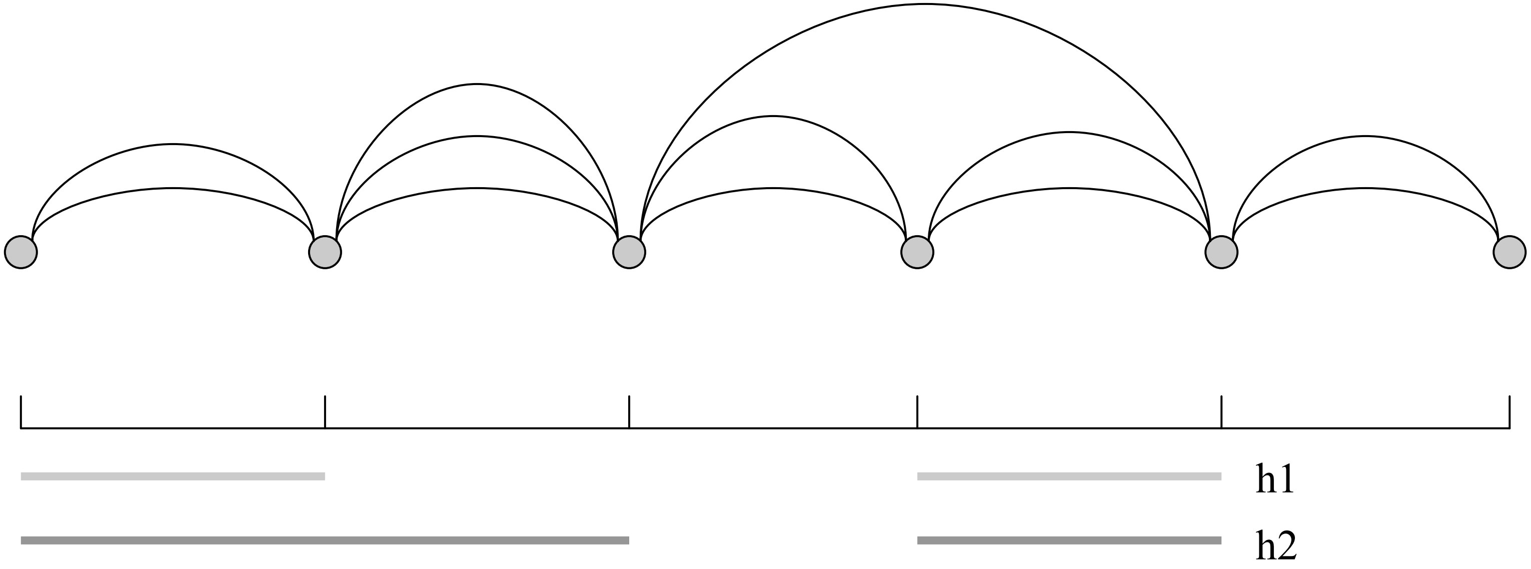
- All permutations
  - Any re-ordering possible
  - Complexity of traveling salesman -> only possible for very short sentences
- Small jumps ahead – filling in the gaps pretty soon
  - Only local word reordering
- Leaving small number of gaps – fill in at any time
  - Allows for global but limited reordering in one direction
  - Similar decoding complexity – exponential in number of gaps
  - IBM-style reordering (described in IBM patent)
- Merging neighboring regions with swap – no gaps at all
  - Allows for global reordering in both directions
  - Complexity lower than 1, but higher than 2 and 3

Dealing with Word Order – Coverage Info

- Need to know which source words have already been translated
  - Don't want to miss some words
  - Don't want to translate words twice
  - Can compare hypotheses which cover the same words
- Coverage vector to store this information
  - For 'small jumps ahead': position of first gap plus short bit vector
  - For 'small number of gaps': array of positions of uncovered words
  - For 'merging neighboring regions': left and right position
- All simple and easy for sentences as input – but does this work for lattices?

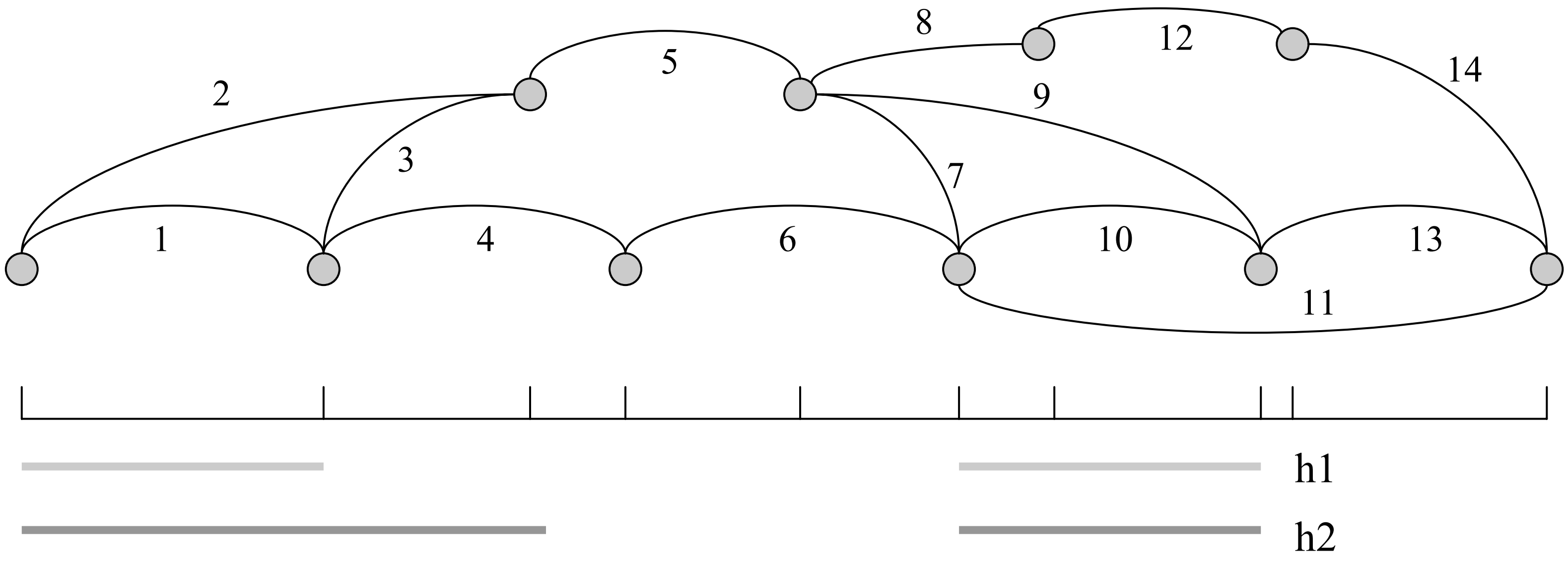
Coverage Information – Confusion Nets

- Empty edges can be eliminated by expanding all incoming edges to next node
- Structure of confusion net is same as structure of translation lattice after inserting word and phrase translations
- Works fine for all reordering strategies



Coverage Information – General Lattices

- Reordering strategy 1 is somewhat messy
  - Can be simulated with strategy 2, so don't bother
- Strategies 2 and 3 pose no additional problem
  - For 2: Recombine hyps when ending in same node and having same untranslated edges
  - For 3: Recombine hyps which have same left and right node



Using Source Language Model

- Edges for target words/phrases need to be linked to source word edges
- Easiest in reordering strategy 3
  - Run over all words attached to source edges, perhaps
  - Run over all words in target edges
  - One side could have the two joint segments swapped
- In reordering strategy 2
  - Calculate source LM prob when leaving word/phrase untranslated
  - Calculate target LM prob when filling the gap

Lattice Translation Results

- A number of papers
  - General word graphs
  - Confusion nets
  - N-best lists (with N=10…100) converted into lattices
- Improvements (small and smaller) reported

- General Picture?

  … there is none

  - Worked last time year - not this year (Wade Shen IWSLT 05 and 06)
  - Works only on small lattices - works on large lattices (Saleem 04, Matusov 07)
  - Works only for mid-range WER – works best for low WER (but your WER ranges might be different from mine)
  - Nice improvements - only small improvements
  - Translation system selects path with lower WER – selects paths with higher WER (Lane, Paulik, inhouse results)

- Problem: our models are too weak to select the few better paths in a word lattice among the zillions of worse paths

Segmentation

- Wrong segmentation introduces errors
  - Reordering over sentence boundaries
  - Loosing LM context
  - Loosing phrase matches
- Segmentation approaches
  - Simply use pause longer then fixed threshold
  - Add LM
  - Add prosodic features
- General picture
  - Wrong segmentation hurts (our TC-Star eval disaster)
  - Longer or shorter segments better?
    - Conflicting results
    - Longer seems better
  - Looking at phrase table seems to help

Punctuation Recovery

Different approaches

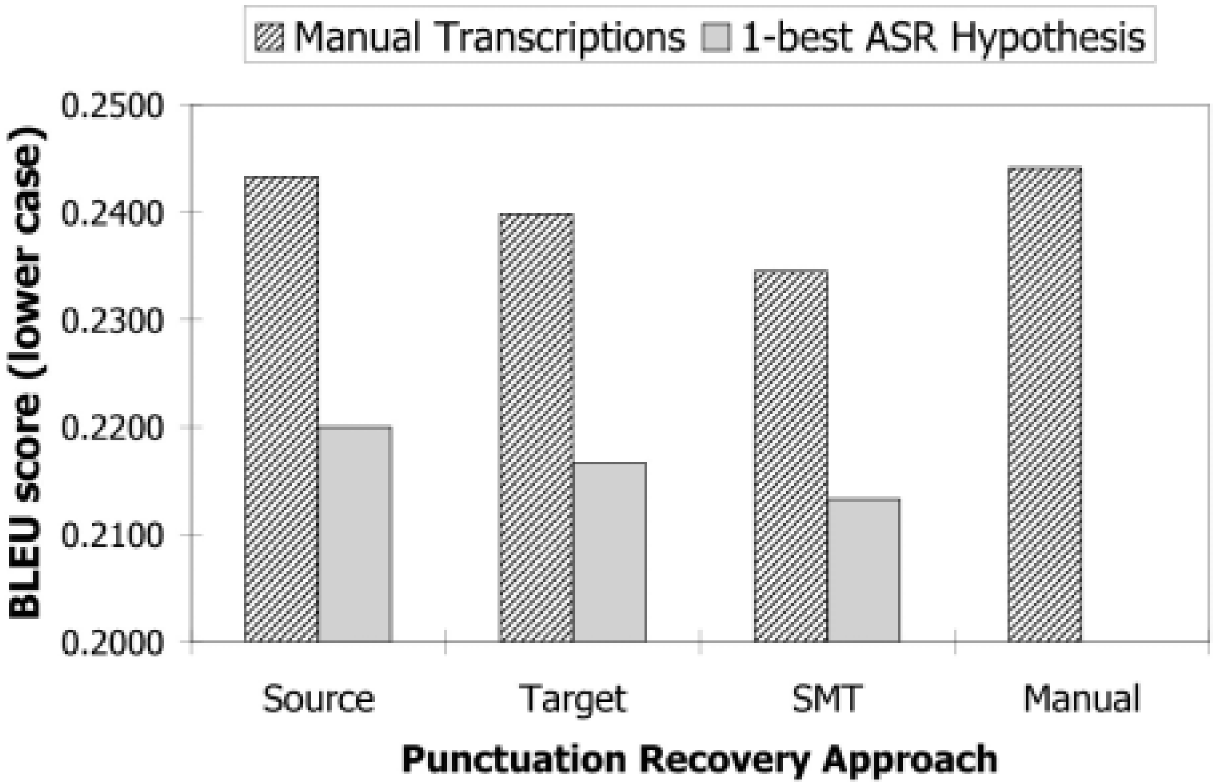
- Use prosodic features as punctuation substitute
- Hidden event LM
  - Delete all punctuation from MT training corpus, restore punctuation on target side through HELM
  - Restore punctuation on source side and used MT trained on corpus with punctuation
- Asymmetric translation model
  - Remove punctuation on source side of training data
  - Or from source side of phrase table
  - Generate punctuation on target side

Punctuation Recovery: Some Results

- IWSLT 2007, Japanese-English translation system
- Recover on source side with HELM

	Precision	Recall	F-score
Manual transcripts			
Sentence Boundary	97.8%	96.8%	97.3%
Secondary Punctuation	82.1%	44.2%	57.5%
1-best ASR Hypothesis: *Character Error Rate=10.4%*			
Sentence Boundary	96.4%	95.9%	96.2%
Secondary Punctuation	71.8%	43.6%	54.3%
Lower Bound: *assume sentence boundary at end of utterance*			
Sentence Boundary	100%	63.9%	77.9%

Punctuation Recovery: Some Results



Source: Punctuation recovery applied to input source
Target: Punctuation recovery applied after translation
(secondary-punctuation ignored during translation)
SMT: Punctuation recovered during translation
Manual: Input source manually annotated with punctuation

User Problems

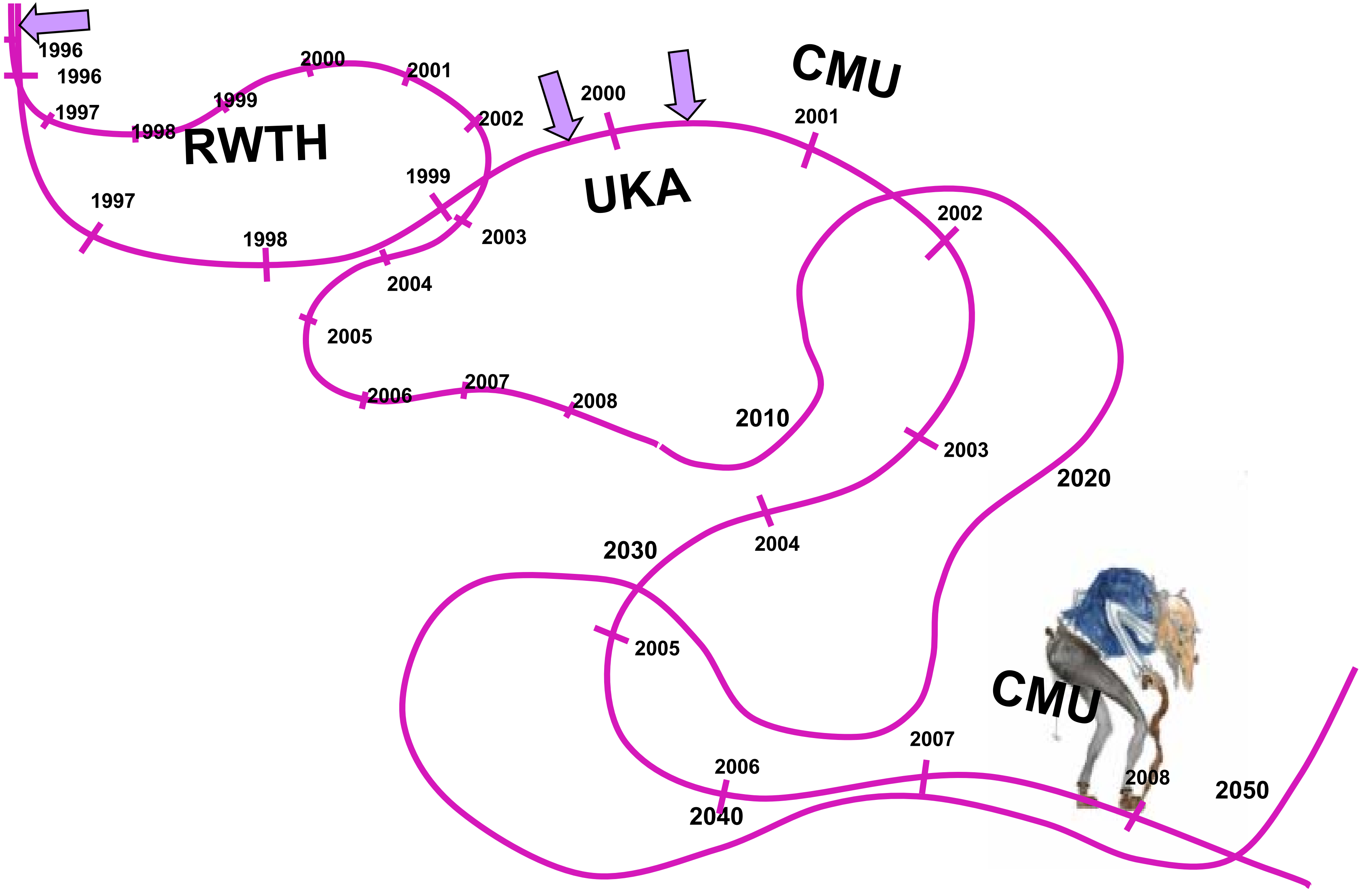
- Output of speech translation systems is difficult to understand
  - Translation quality
  - Missing or incorrect segmentation: Really big problem for user
- Text display
  - Difficult to ‘parse’ the system output
  - Text display without punctuation is difficult to read
  - Wrong line-breaks have to be (mentally) undone
- Speech synthesis
  - Rather monotone and unstructured
  - Overlay with original speech
  - Real-time processing by user is hardly possible

Next Grand SLT Challenge: Multilingual Meetings

- Several meeting projects: AMI and AMIDA, CHIL, ICSI, …
- So far not translation involved, but would be a natural extension
  - Participants have often different native languages
  - There could be a real need, even more than for lecture translation

- Interesting research issues
  - Latency in translation
    - How to achieve short latency
    - How does is affect translation quality
    - What is ‘best’ latency for user
  - How to integrate external knowledge sources
    - Slides, notes from previous meetings
    - Online search for model adaptation
  - Learning from human interpreters
    - Strategies to translate despite of missing information
  - Presenting the translations
    - Audio or video display
    - How to deal with cross talk and parallel discussions
    - Tele- and videoconferences seem to provide an easier setup

My Personal (Speech) Translation Journey ???


1996
1996
1997
1998
1999
2000
2001
2002
RWTH
1997
1998
1999
2003
2004
2005
2006
2007
2008
2010
2000
UKA
CMU
2001
2002
2003
2004
2005
2006
2007
2008
2020
2030
2040
2050
CMU