Machine Translation
Introduction

Francis Bond
NTT Communication Science Laboratories
www.kecl.ntt.co.jp
bond@cslab.kecl.ntt.co.jp

2006-07-10: lecture 1

Course Outline

(1) **Introduction**

- Why do Machine Translation?
- Approaches to Machine Translation
  - Rule-based (Knowledge-based): Transfer, Interlingual
  - Example-based: Statistical, Case-based, Translation Memories
  - Combinations: Hybrid, Multi-engine

(2) Case Studies

- An in depth-look at some MT Systems
  - Analysis and Generation
  - Transfer
  - Tuning and Adaptation
- Conclusion and References

Outline for Lecture 1

- Outline
- The demand for Machine Translation
- Problems
  - Linguistic
  - Technical
  - Interface
- Kinds of Machine Translation
  - Rule-based (Knowledge-based): Transfer, Interlingual
  - Example-based: Statistical, Case-based
  - Combinations: Hybrid, Multi-engine
- Successful and Unsuccesful Applications
- The Future

Increased Demand

- Growing amount of cross-lingual communication
  - A tenth of the U.N. Budget
  - Over €1,000,000,000 for the EU every year
  - Global Economy
  - Easy access over the internet
    - Google Translation is their most used special feature
- Large amounts of machine readable text
  - Increase in the use of computers
  - Improvement of scanners and speech-to-text systems
- A desire for quick translation

Linguistic Background

- No settled linguistic theory$_i$ exists
  - Can't just implement it$_i$
  - Non-core phenomena are very common
    often neglected by mainstream linguist research
- Translation is AI complete.
  - Requires full knowledge of the world.
  - Often requires specialist domain knowledge
  - Even humans make mistakes

Parsing

- What should the output be for *I like words* ?
  - syntactic trees?
    (S (NP I) (VP (V like) (NP (N words))))
  - semantic logical forms?
    [`like(speaker,word+PL)`]
  - pragmatic speech acts?
    Speaker wants hearer to believe that speaker believes that `like(speaker,word+PL)`]
  - whatever is useful?
    watashi-wa kotoba-ga suki-da

- How to model an infinite set of expressions?

- What should the basic units of translation be?

Transfer — equivalents?

- Category changes: *postwar* $_{adj} \rightarrow$ *nach dem Krieg* $_{np}$
- Lexical gaps: *wear* $\rightarrow$ *haku* “wear below waist”, *kiru* “wear above waist”, *kaburu* “wear on head”
- Head switching:

  (1) *I swam across the river*

  (2) *J’ai traversé le fleuve en nageant*

  I crossed the river by swimming

Transfer — mismatches

"The differences in languages lie not in what you can say, but rather what you must"
Roman Jakobson

- number
- definiteness
- gender
- politeness
- evidentiality

Transfer — discourse

- Different discourse order in Japanese and American stockmarket reports
- Differing conventional implicatures
  -te-mo ii “conditional” is much less positive than *you may*
- *Must you go, can’t you stay?* (in middle class English)
  bubu-duke ikaga-desuka “would you like some rice and tea” (Kyoto)
  $\Rightarrow$ `go home at once!`
- Some work on speech acts in the Verbmobil project
- ⊗ All to often ignored entirely

Technical Limitations

- Problems of Economy
  - Memory Limitations
  - Speed Problems
    Some recent improvements in parallel processing
- Problems of Consistency
  - Increased lexical choice leads to less consistency
  - Large systems are often hard to predict
- The need for more information

Knowledge Acquisition

- Unknown words:
  Yahoo, *sidewalk*, *togs*
- Unknown senses:
  (satellite) *footprint*, (system) *daemon*
- Unknown relationships:
  Machine translation is easy, *NOT!*
- Partially solved by:
  - Domain Specific Lexicons (and rules)
    Terminology
  - Register Specific Lexicons (and rules)
  - Knowledge Acquisition from Corpora

Interfaces

- OCR
- Speech-to-Text
  - Almost always impoverished
    no prosody, no spelling, no Chinese characters
  - Is frequently wrong
    wreck a nice peach vs *recognize speech*
- Text
  - Must often be cleaned
    correct spelling errors, loose fancy fonts
  - May have useful structural mark up
    `list header`, `list item`

Various approaches to MT

- Rule-based: RBMT (transfer-based, knowledge-based)
- Example-based: EBMT
- Statistical: SMT

Rule-based MT

- Parse SL to some more abstract form: the meaning?

  The dog chases a cat
  → chase$_1$(dog$_1$:[def], cat$_1$:[indef])

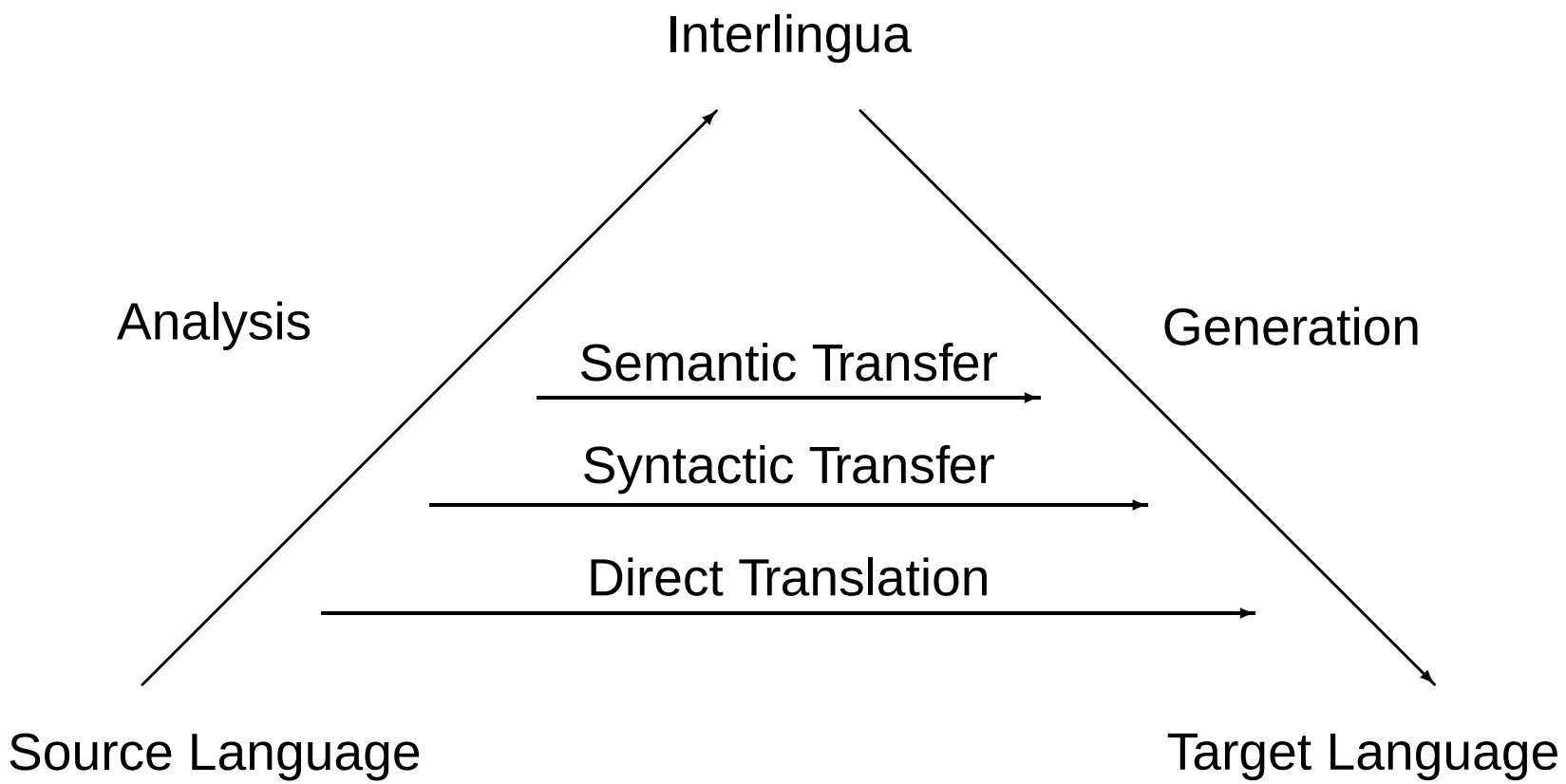
- Transfer to the target language abstract form

  → 追う$_1$(犬$_1$:[def], 猫$_1$:[indef])

- Generate from this

  → 犬 が 猫 を 追う
  inu ga neko wo ou
  dog NOM cat ACC chase

The Vauquois Triangle

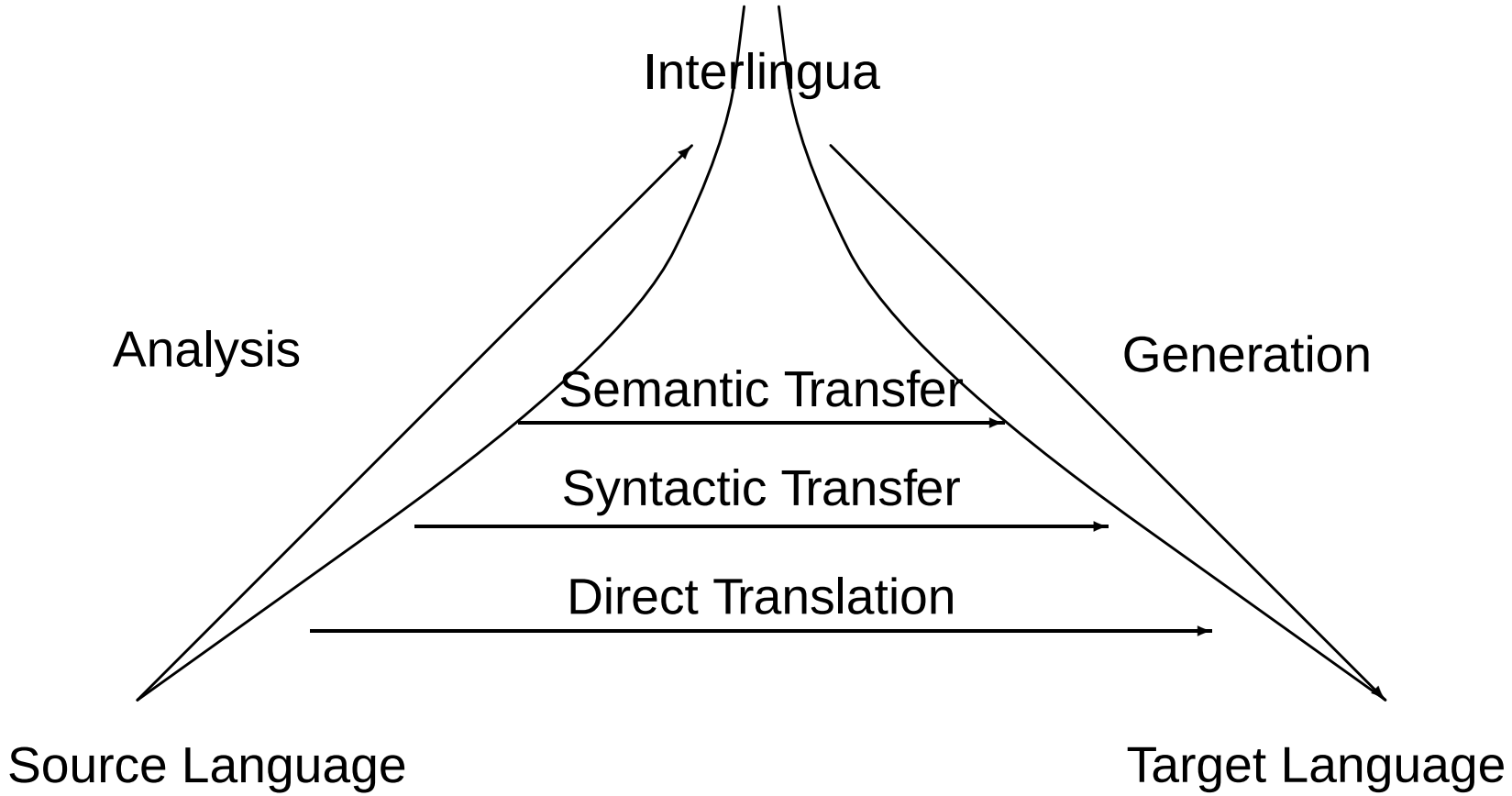


Transfer vs Interlingua

- Transfer Based: $n(n-1)$
  - Commercial systems: SYSTRAN, METAL, L&H etc
  - Research systems: ALT-J/E, Verbmobil, Logon, OpenTrad
- Interlingua: $2n$
  - Multilingual systems: Eurotra, CCIC, UNL

There is a convergence in real life.

The Ikehara Discontinuity



RBMT: Summary

- This is the classic approach in NLP
- RBMT is the most widely used commercially
  - Many existing systems
  - Can customize, mainly by adding/removing words to the lexicons
- RBMT suffers from the knowledge acquisition bottleneck
  - Building lexicons is expensive (2-20 AUD/word)
  - It is hard to set defaults by hand
  - Rule interactions are hard to understand in a big system

Example based MT

- Case Based:
  - Kyoto University: Nagao et al.
  - ATR: TDMT
  - Dublin University
- Memory-based translation:
  - Translation Memories
    Very popular as an aid

EBMT Basic Philosophy

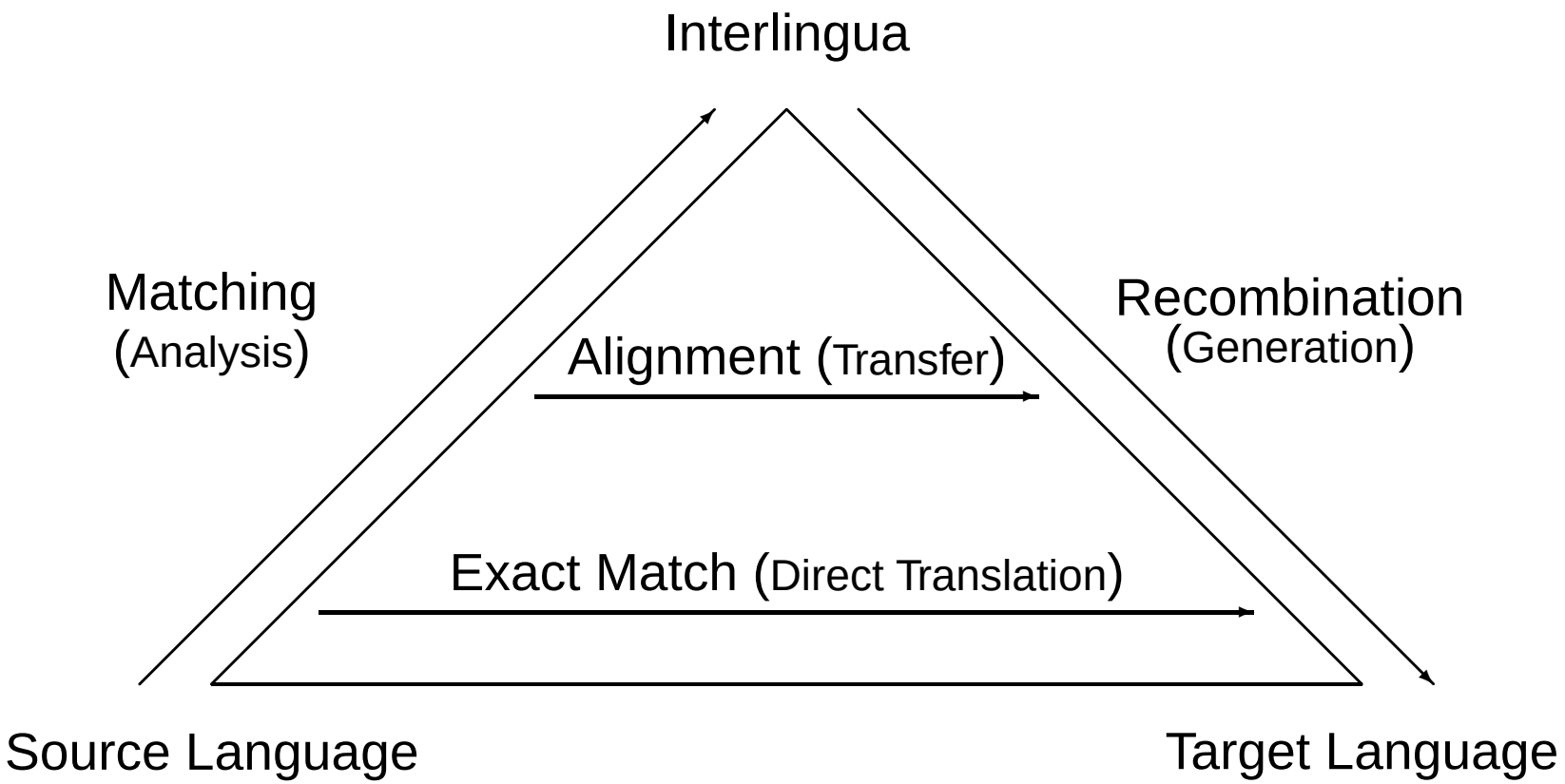
"Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference."

Makoto Nagao (1984)

EBMT philosophy

- When translating, reuse existing knowledge:
  - Match input to a database of translation examples
  - Identify corresponding translation fragments
  - Recombine fragments into target text
- Example:
  - Input: He buys a book on international politics
  - Data:
    - He buys a notebook – Kare wa noto o kau
    - I read a book on international politics – Watashi wa kokusai seiji nitsuite kakareta hon o yomu
  - Output: Kare wa kokusai seiji nitsuite kakareta hon o kau

EBMT ‘Pyramid’



H. Somers, 2003, “An Overview of EBMT,” in *Recent Advances in Example-Based Machine Translation* (ed. M. Carl, A. Way), Kluwer

Example-based Translation: Advantages/Disadvantages

- Advantages
  - Correspondences can be found from raw data
  - Examples give well structured output
- Disadvantages
  - Lack of well aligned bitexts
  - Generated text tends to be incohesive

State of the Art

- EBMT does best with well aligned data in a narrow domain
  - There are not so many domains with such data
- EBMT not used in commercial systems
- EBMT eclipsed by SMT in competitions
- Still a healthy research community
- EBMT and SMT converging
  - EBMT adds probablisitic models
  - SMT adds larger phrases

Translation Memories (1)

- **Translation Memories** are aids for human translators
  - Store and index existing translations
  - Before translating new text
    - Check to see if you have translated it before
    - If so, reuse the original translation
- Checks tend to be very strict $\Rightarrow$ translation is reliable
  - Identical except for white-space differences
- Now extended to **fuzzy** matching and replacing
  - Equivalent to EBMT
  - More flexible, greater cover, less reliable

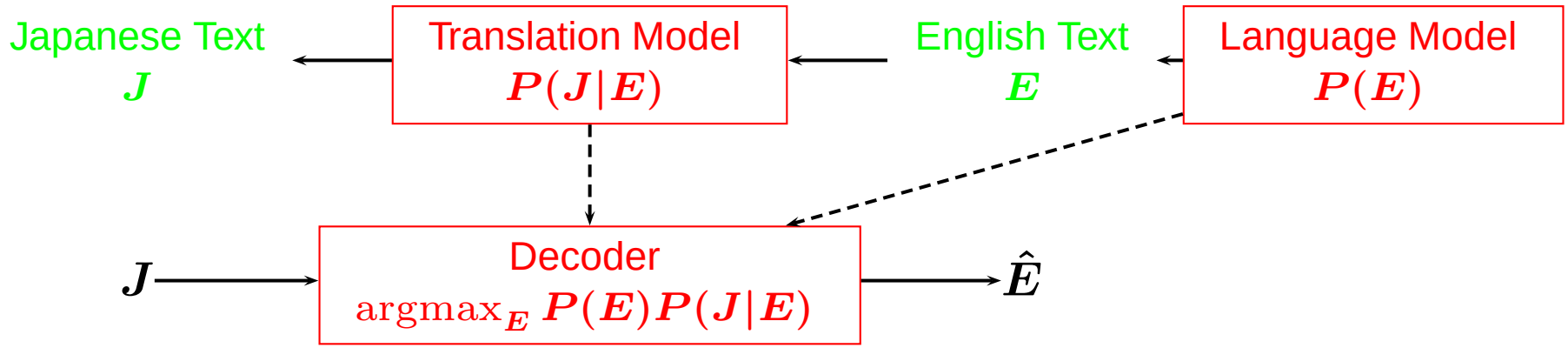
Translation Memories (2)

- TM is popular with translators
- Well integrated with word processors
- The translator is in control
- Translation companies can pool memories, giving them an advantage
- Simple solutions sell well

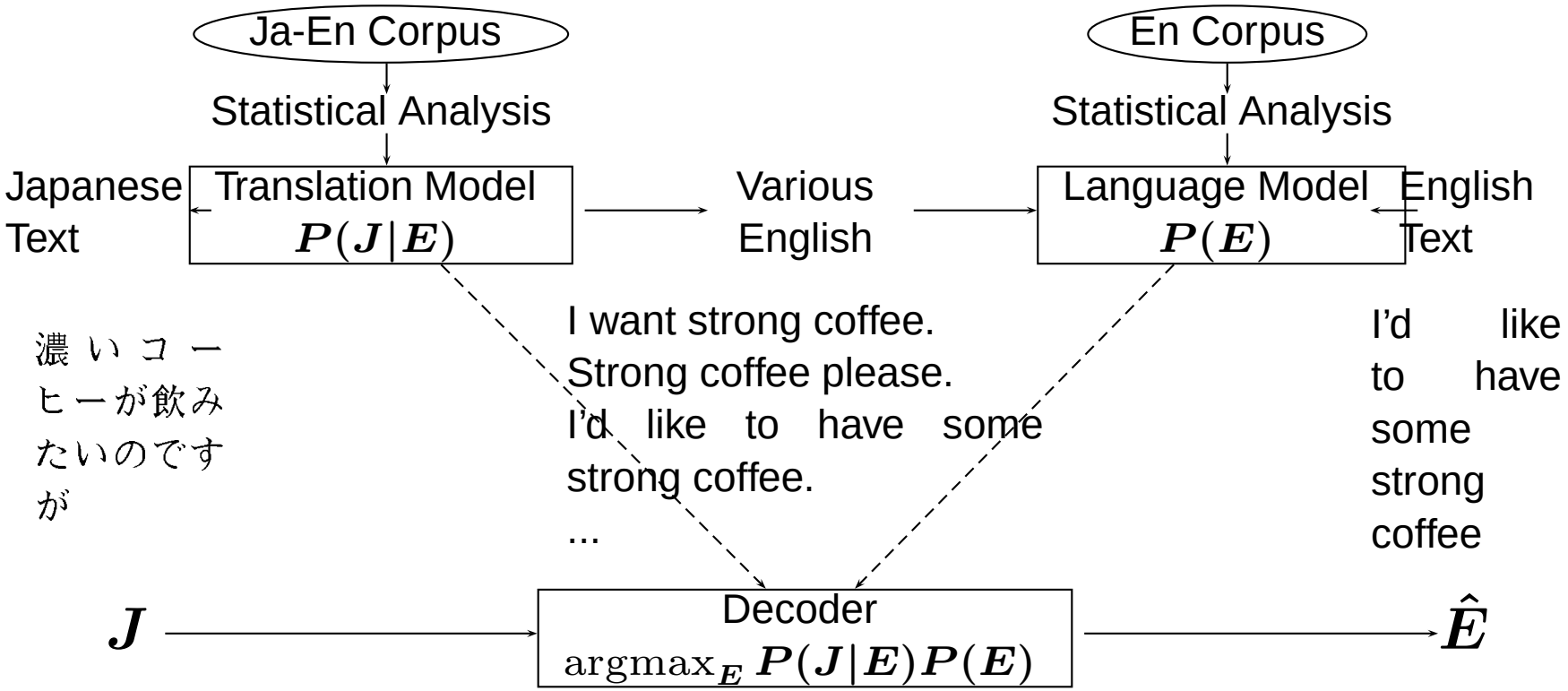
Statistical Machine Translation

- The Basic Idea (Brown *et al* 1990)
  Find the most probable English sentence given a foreign language sentence

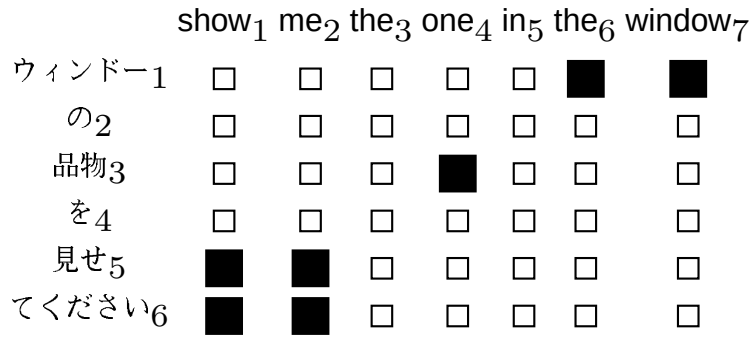
$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|J)$$



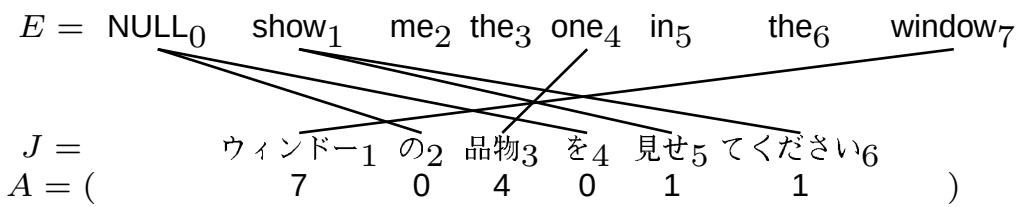
Statistical MT Framework



Aligning Text



- A compact Representation



- How many possible alignments? $\longrightarrow (l+1)^m$

The Translation Model (IBM Model 4)

$$P(J, A|E)$$

Fertility Model

$$\prod n(\phi_i|E_i)$$

NULL Generation Model

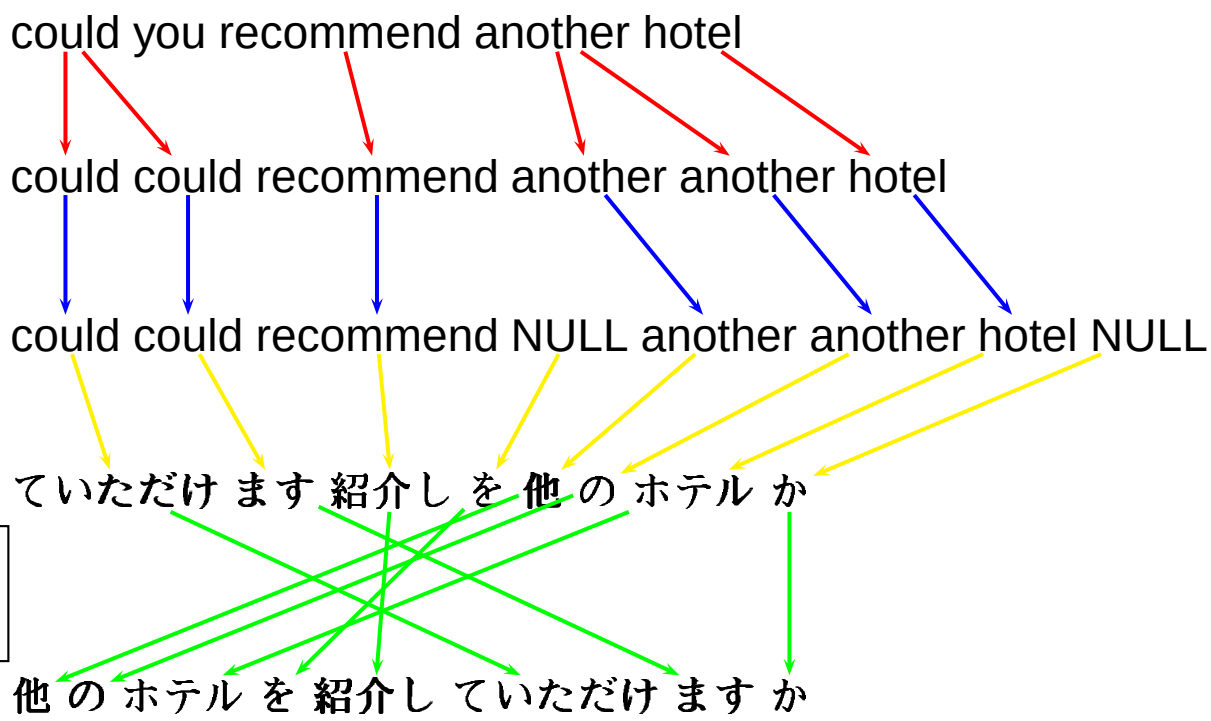
$$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$$

Lexicon Model

$$\prod t(J_j|E_{A_j})$$

Distortion Model

$$\prod d_1(j - k|\mathcal{A}(E_i)\mathcal{B}(J_j))$$
$$\prod d_{1>}(j - j'|\mathcal{B}(J_j))$$



Current Problems

- Translation of long sentences
  - Complex sentences and coordination
- Corpus size
  - Is more always better?
  - Do errors in the corpus matter?
- Efficiency
  - The current best system takes one hour/sentence
- Unknown words

SMT:Summary

- Currently the hottest area of research
- Commercial systems just being deployed (En-Ar, En-Cn)
- So far more data trumps more complicated models
  - Doubling the translation model $\Rightarrow$ 2.5 increase in BLEU score
  - Doubling the language model $\Rightarrow$ 0.5 increase in BLEU score
- Still a lot of research on more complicated models
  - You can't always get twice as much data
  - It is hard to customize systems

MT Evaluation: The BLEU score

- Evaluating MT output is non-trivial
  - There may be multiple correct answers.
    - *I like to swim*, *I like swimming*
    - *Swimming turns me on*
- Hand evaluation requires a bilingual evaluator - expensive
- Automatic evaluation can be done by comparing results (in a held out test set) to a set of reference translations
  - The most common metric is BLEU
    compares n-gram overlap with a brevity penalty
  - 0.3–0.5 typical; 0.6+ approaches human
  - Correlates with human judgement, but not exactly
  - Other score are Word Error Rate; NIST (weighted BLEU)

Combinations

- Multiengine:
  - CMU/ISI: Pangloss
- Hybrid:
  - NTT: Hybrid-ALT
- Dialogue-based:
  - GETA: Interactive Disambiguation (Lydia)
    The AD/ID/AD Sandwich

Successful Applications

Controlled language

- Narrow Domain:
  - Canada: Meteo
  - NTT: ALTFLASH
- Controlled Language:
  - CMU: KANT
    Control languages

Browsing

- Security summaries
  - the original aim!
- Internet access
  - SYSTRAN, Pensee, Babelfish and many more

Machine Aided Translation

- Translation memory
- Dictionary look up/construction
- Automatic glossing
- Writing Assistance

Unsuccessful Applications

Fully Automatic High Quality Translation

But we still keep trying . . .

Spinoffs

- Automatic Proof-reading
- Writing Assistants
  Spell Checkers, Grammar Checkers
- Text-to-Speech
- Text-to-Braille
- Hand held lexicons