Rule-Based MT, including Knowledge-Based MT (KBMT)

11-731 MT and Seq2Seq Models

Robert Frederking

31 January 2017

MT History

- Did not start with IBM Model I Statistical MT

MT history

- 1933 – Patents in France and Russia for mechanical translation devices
- 1940s – WW2 code breaking efforts
- 1947 – Weaver letter outlining translation as a problem in cryptography
- 1954 – Georgetown Experiments showed "promise" of Russian-English MT
- 1966 – ALPAC report shifts funding to basic research in computational linguistics
- 1968 – MT company SYSTRAN founded (still in existence)
- 1970s – advances in formal languages and automata theory; development of *statistical speech recognition* techniques at IBM and Princeton and CMU
- **1985 – CMU's Center for Machine Translation founded**
- **1980s – Domain-specific MT developed, Speech-to-speech MT begun**
- 1993 – Weaver's model of translation prototyped by IBM; *statistical revolution*
- **1996 – Center for Machine Translation becomes LTI**
- 1999 – Open source reimplementation of IBM statistical models
- 2000s – Major modeling improvements, rediscovery of syntax, large scale funding
- 2006 – Google Translate launches
- 2010 – SDL (translation company) acquires Language Weaver (MT company)

State-of-the-Art in MT:

- What users really want:
  - General purpose (any text)
  - High quality (human level)
  - Fully automatic (no user intervention)
- We can meet any 2 of these 3 goals today, but not all three at once!

State-of-the-Art in MT:

- What users want:
  - General purpose (any text)
  - High quality (human level)
  - Fully automatic (no user intervention)
- We can meet any 2 of these 3 goals today, but not all three at once:
  - FA HQ: Knowledge-Based MT (KBMT)
  - FA GP: Corpus-Based (SMT/EBMT) MT
  - GP HQ: Human-in-the-loop (efficiency tool)

Central Problems of MT:

- **Ambiguity:**
  - Human languages are highly ambiguous, and differently in different languages.
- **Amount of knowledge:**
  - At least several 100k words, about as many phrases, plus syntactic knowledge. **How** do you make a knowledgebase that big that is (even mostly) correct and consistent?
- Syntactic complexity not as big an issue!

MT: math or application?

- Research funding is now almost all SMT or NN
- If your interest is MT as a real-world application, many other issues come up:
  - Application types
  - Human translators
  - Human factors
  - User support
  - etc…

Types of MT Applications:

- **Assimilation**: multiple source languages, uncontrolled style/topic. General purpose MT, no semantic analysis. (GP FA or GP HQ)
- **Dissemination**: one source language, controlled style, single topic/domain. Special purpose MT, full semantic analysis. (FA HQ)
- **Communication**: Lower quality may be okay, but degraded input, real-time required.

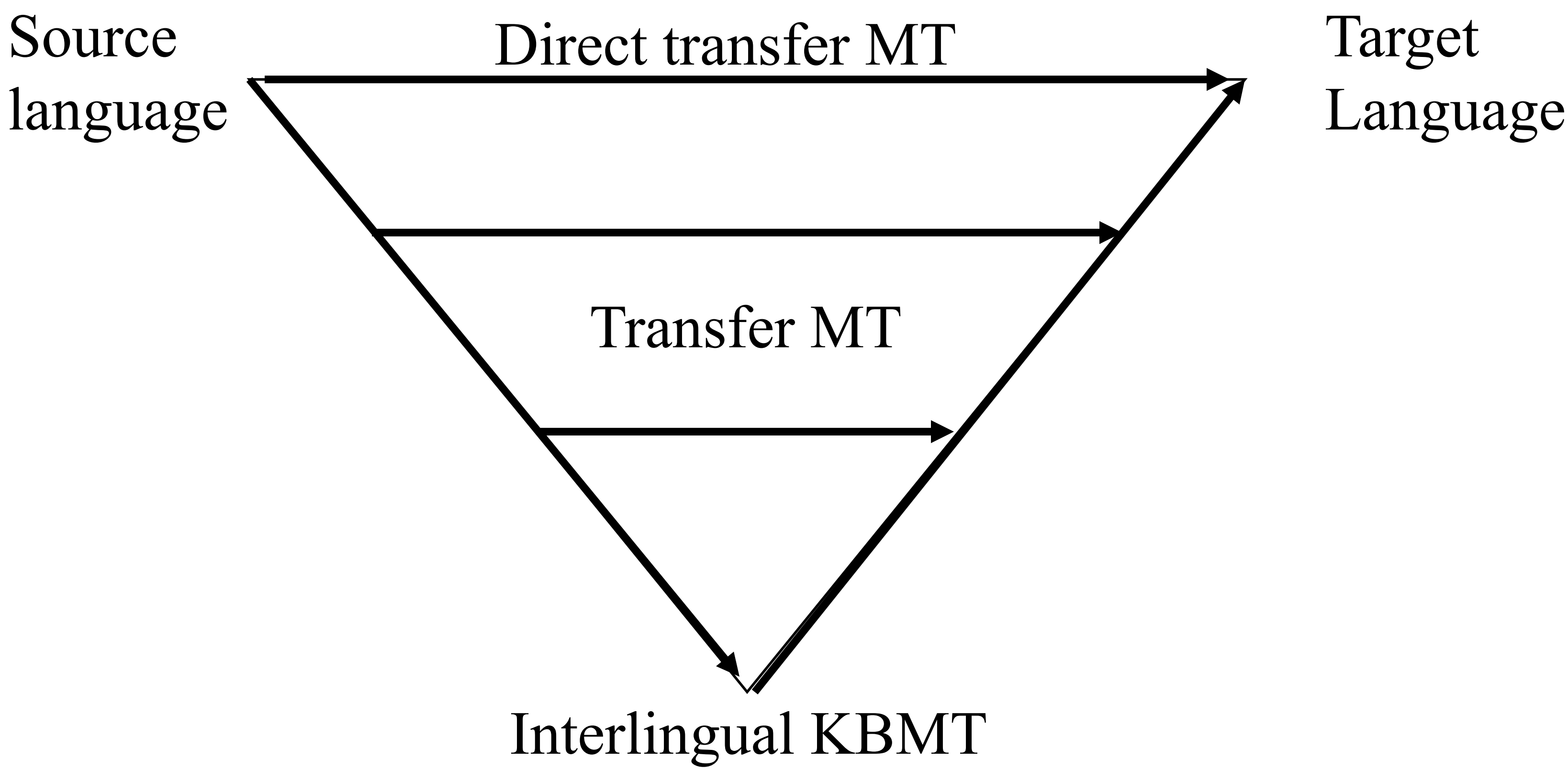
LTI’s MT History

- High-Accuracy Interlingual MT
  - KANT: large-scale, practical MT for technical documentation
    - First high-accuracy text MT
- Speech-to-speech MT
  - JANUS/Nespole!/LingWear/DIPLOMAT/Tongues/Babylon/ TransTac:
    - First speech-speech MT (JANUS)
    - Jibbigo bought by Facebook
- Parallel Corpus-Trainable MT
  - Statistical MT
  - Example-Based MT (→ Phrase-Based SMT)

LTI's MT History (cont.)

- Multi-Engine MT: first MT ensemble approach
- METEOR MT metric: best fit to human judges
- MT-related systems:
  - First high-accuracy translingual IR
- Endangered Language MT:
  - First minority-language MT (DIPLOMAT)
  - AVENUE, …, LORELEI
- Spin-off companies: (besides Jibbigo)
  - Safaba bought by Amazon (now Amazon Pgh!)

Types of MT technologies



[After Vauquois]

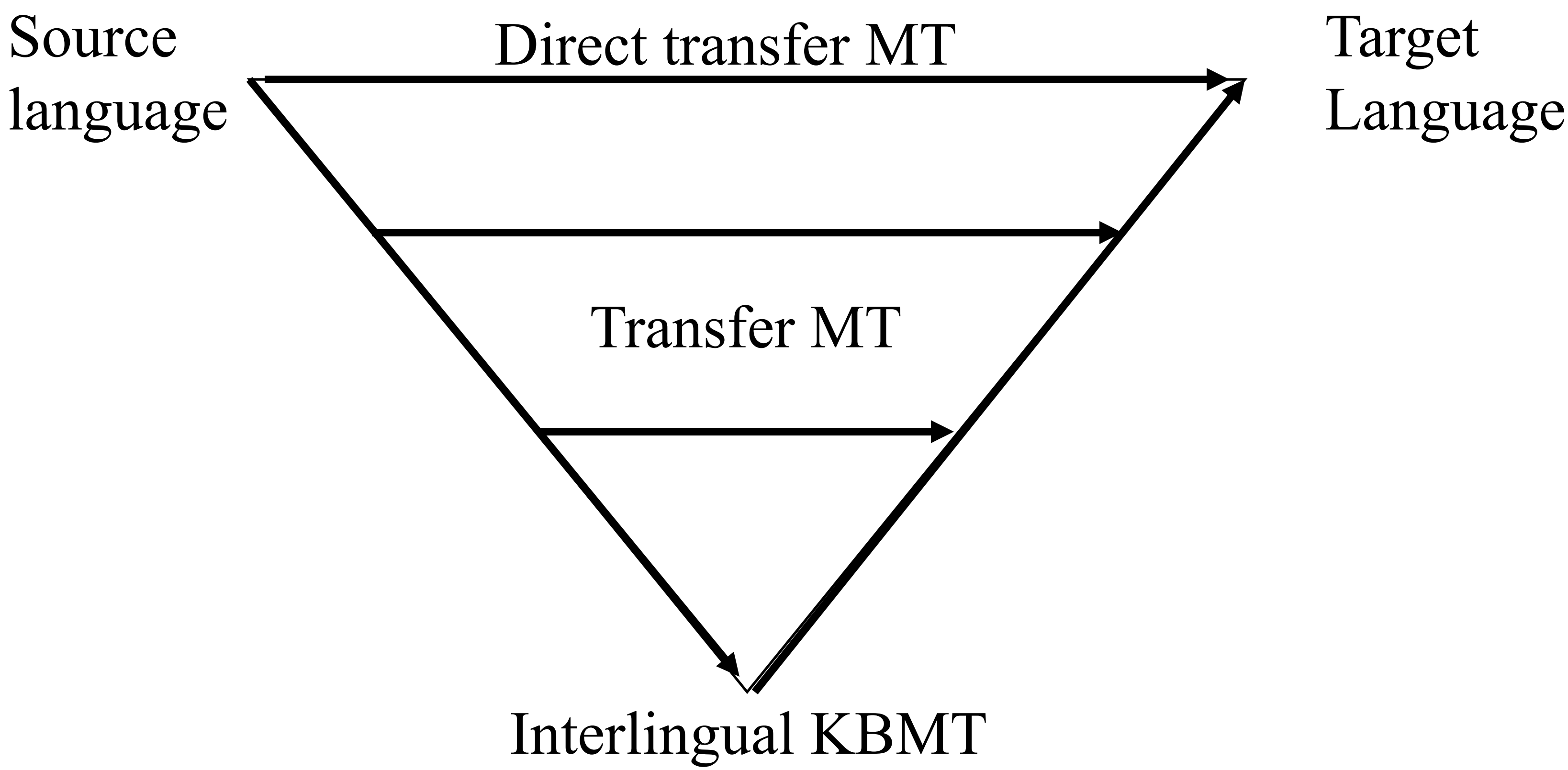
Direct transfer MT

- Earliest approach to MT
- *Huge* dictionaries of bilingual phrase pairs
- Heuristics to pick among ambiguous choices
- Could also add semantic fields, kitchen sink

- But lots of tricky cases:

  I like to swim → Ich schwimme gern

Types of MT technologies

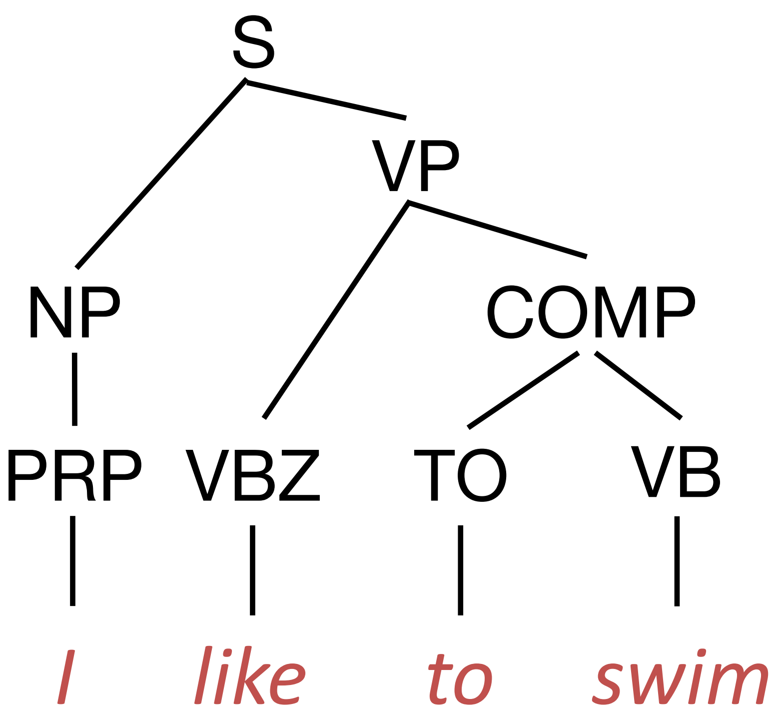


[After Vauquois]

Syntactic Transfer MT

- Basic idea:
  - analyze a Source Language sentence to get the *syntactic structure*,
  - apply transfer rules to convert SL syntax into TL syntax
  - then generate a Target Language sentence that respects TL syntactic constraints, inserting TL lexical items

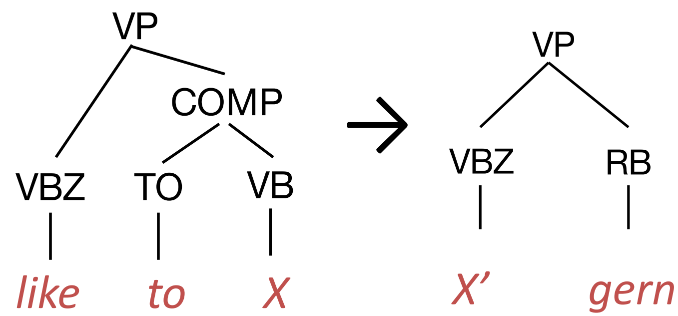
Syntactic Transfer MT



Transfer rules:

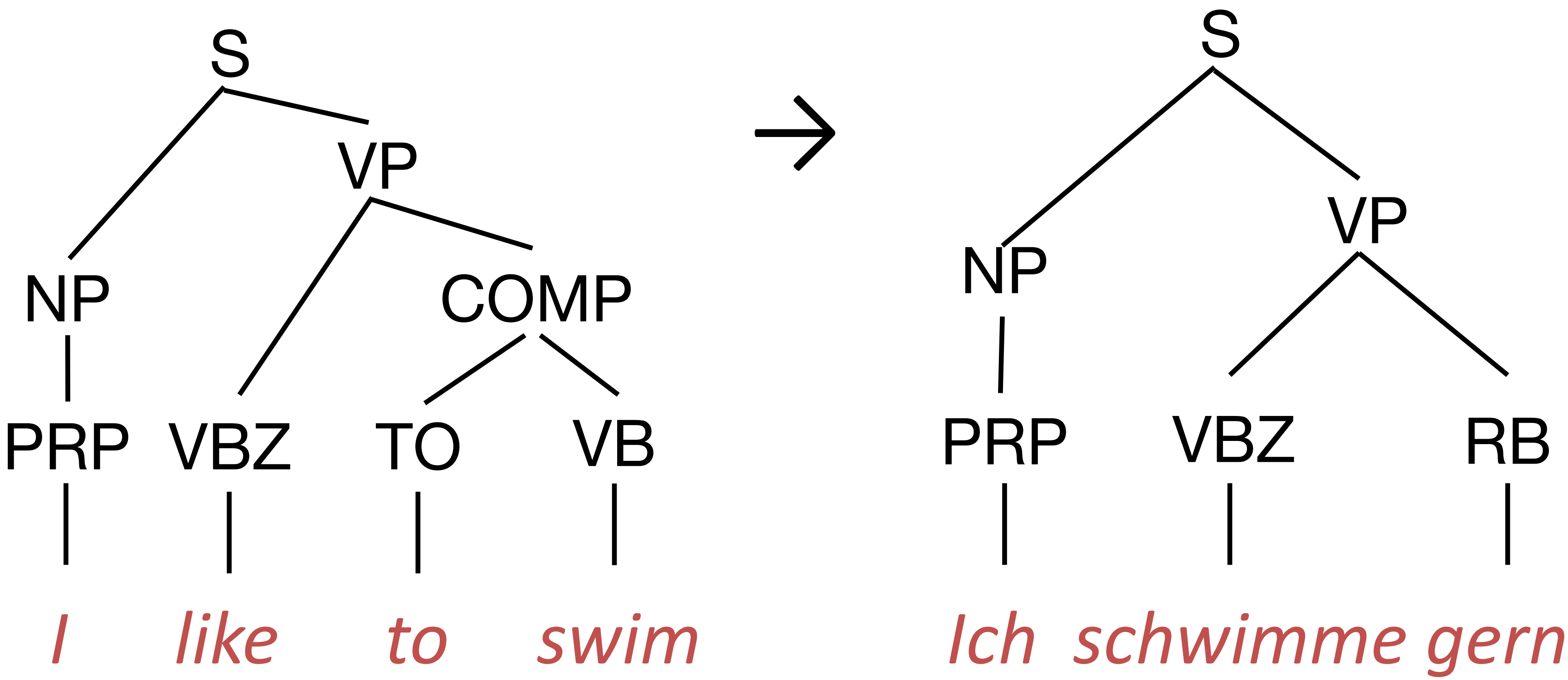
I → ich

swim → schwimme

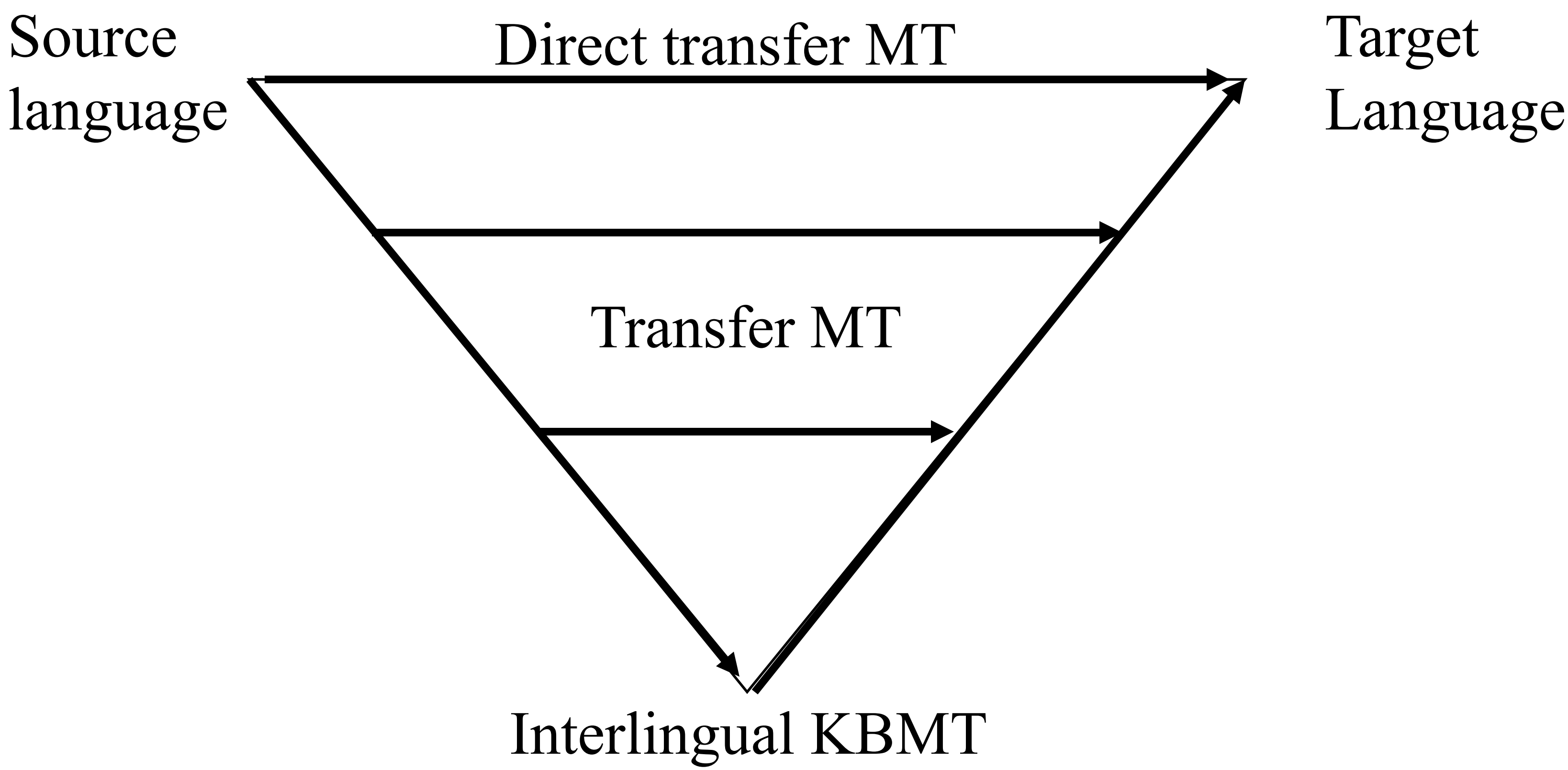


I like to swim → Ich schwimme gern

Syntactic Transfer MT



Types of MT technologies



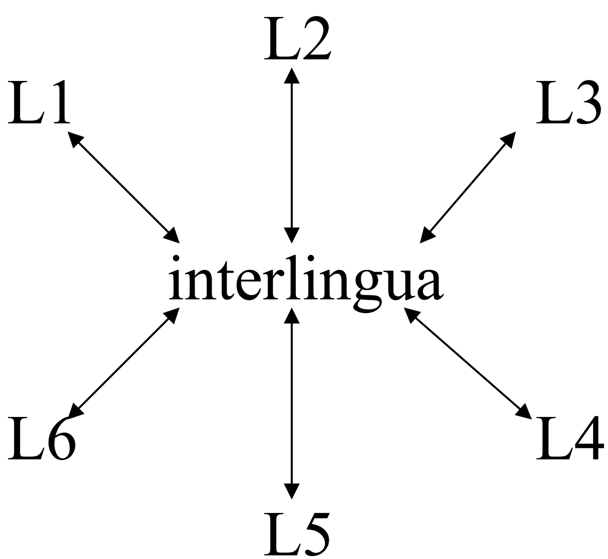
[After Vauquois]

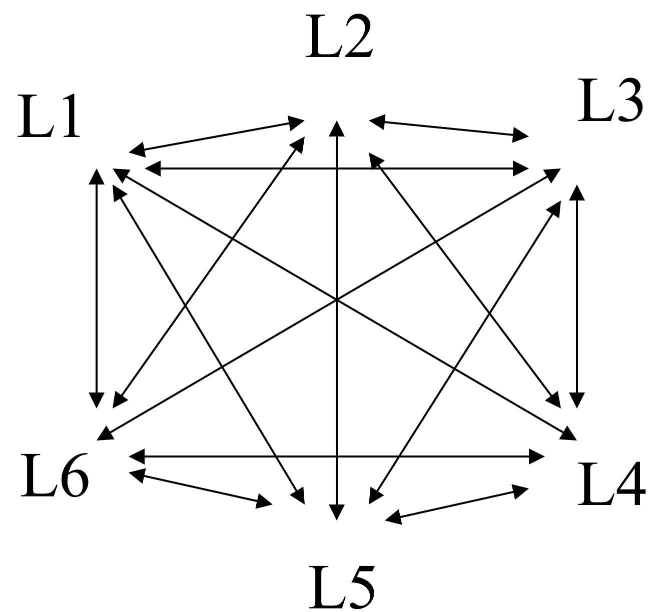
Knowledge-based Interlingual MT

- The "obvious" Artificial Intelligence approach to MT:
  - Analyze the input language to find the meaning
  - Generate this meaning in the output language
- "Interlingual": one meaning representation for all languages
  - May or may not be possible in general!

The Interlingua KBMT approach:

- With interlingua, need only N parsers/ generators instead of N^2 transfer systems:





Knowledge-Based MT (KBMT)

- Basic KBMT idea:
  - analyze a Source Language sentence to get the *meaning*,
  - then generate a Target Language sentence that expresses that *meaning*

- Hmm: so how do you represent *sentence meanings*?

Representing NL meaning

- Fortunately, there has been a lot of work on this (since Aristotle, at least)
  - Panini in India too
- Especially, formal mathematical logic since 1850s (!), starting with George Boole etc.
  - Wanted to replace NL proofs with something more formal
- Deep connections to set theory

Model-Theoretic Semantics

- Model: a simplified representation of (some part of) the world: **sets** of objects, properties, relations (**domain**).
- Logical vocabulary: like reserved words in PL
- Non-logical vocabulary
  - Each element **denotes** (maps to) a well-defined part of the model
  - Such a mapping is called an **interpretation**

A Model

- **Domain**: Noah, Karen, Rebecca, Frederick, Green Mango, Casbah, Udipi, Thai, Mediterranean, Indian
- **Properties**: Green Mango and Udipi are crowded; Casbah is expensive
- **Relations**: Karen likes Green Mango, Frederick likes Casbah, everyone likes Udipi, Green Mango serves Thai, Casbah serves Mediterranean, and Udipi serves Indian

- n, k, r, f, g, c, u, t, m, i
- Crowded = {g, u}
- Expensive = {c}
- Likes = {(k, g), (f, c), (n, u), (k, u), (r, u), (f, u)}
- Serves = {(g, t), (c, m), (u, i)}

Some English

- *Karen likes Green Mango and Frederick likes Casbah.*
- *Noah and Rebecca like the same restaurants.*
- *Noah likes expensive restaurants.*
- *Not everybody likes Green Mango.*

- What we want is to be able to represent these statements in a way that lets us compare them to our model.
- **Truth-conditional semantics**: need operators and their meanings, given a particular model.

First-Order Logic

- **Terms** refer to elements of the domain: **constants**, **functions**, and **variables**
  - Noah, SpouseOf(Karen), X
- **Predicates** are used to refer to sets and relations; predicate applied to a term is a **Proposition**
  - Expensive(Casbah)
  - Serves(Casbah, Mediterranean)
- Logical connectives (**operators**):

  $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\Rightarrow$ (implies), ...
- **Quantifiers** ...

Quantifiers in FOL

- Two ways to use variables:
  - refer to one anonymous object from the domain (**existential**; $\exists$; "there exists")
  - refer to all objects in the domain (**universal**; $\forall$; "for all")

- *A restaurant near CMU serves Indian food*
  $\exists x \text{ Restaurant}(x) \wedge \text{Near}(x, \text{CMU}) \wedge \text{Serves}(x, \text{Indian})$
- *All expensive restaurants are far from campus*
  $\forall x \text{ Restaurant}(x) \wedge \text{Expensive}(x) \Rightarrow \neg\text{Near}(x, \text{CMU})$

FOL: Meta-theory

- Well-defined set-theoretic semantics
- **Sound:** can’t prove false things
- **Complete:** can prove everything that logically follows from a set of axioms (e.g., with “resolution theorem prover”)
- Well-behaved, well-understood
- Mission accomplished?

FOL: But there are also "Issues"

- "Meanings" of sentences are ***truth values***.
- Only ***first-order*** (no quantifying over *predicates* [which the book does without comment]).
- Not very good for ***"fluents"*** (time-varying things, real-valued quantities, etc.)
- Brittle: ***anything*** follows from *any* contradiction(!)
- ***Goedel incompleteness***: "This statement has no proof"!
  - (Finite axiom sets are incomplete w.r.t. the real world.)

- **So:** Most systems use its descriptive apparatus (with extensions) but not its inference mechanisms.

Extending FOL

- To handle sentences in non-mathematical texts, you need to cope with additional NL phenomena
- Happily, philosophers/logicians have thought about this too

Generalized Quantifiers

- In FOL, we only have universal and existential quantifiers
- One formal extension is type-restriction of the quantified variable: *Everyone likes Udipi*:

  $\forall$ x Person(x) $\Rightarrow$ Likes(x, Udipi) becomes

  $\forall$ x | Person(x).Likes(x, Udipi)
- English and other languages have a much larger set of quantifiers: *all, some, most, many, a few, the, …*
- These have the same *form* as the original FOL quantifiers with type restrictions:

  <quant><var>|<restriction>.<body>

Generalized Quantifier examples

- *Most dogs bark*

  Most x | Dog(x) . Barks(x)

- *Most barking things are dogs*

  Most x | Barks(x) . Dog(x)

- *The dog barks*

  The x | Dog(x) . Barks(x)

- *The happy dog barks*

  The x | (Happy(x) $\wedge$ Dog(x)) . Barks(x)

- ***Interpretation*** and ***inference*** using these are harder…

Semantic Cases/Thematic Roles

- Another aspect of semantics not represented in traditional FOL

- Developed in late 1960's and 1970's
- Postulate a limited set of abstract semantic relationships between a verb & its arguments: *thematic roles* or *case roles*

Thematic Role example

- *John broke the window with the hammer*
- *John*: AGENT role
  window: THEME role
  hammer: INSTRUMENT role
- Extend LF notation to use semantic roles

Can We Generalize?

- **Thematic roles** describe general patterns of participants in generic events.
- This gives us a kind of shallow, partial semantic representation.
- First proposed by Panini, before 400 BC!

Thematic Roles

Role	*Definition*	*Example*
Agent	Volitional causer of the event	**The waiter** spilled the soup.
Force	Non-volitional causer of the event	**The wind** blew the leaves around.
Experiencer		**Mary** has a headache.
Theme	Most directly affected participant	Mary swallowed **the pill**.
Result	End-product of an event	We constructed **a new building.**
Content	Proposition of a propositional event	Mary knows **you hate her**.
Instrument		You shot her with **a pistol.**
Beneficiary		I made **you** a reservation.
Source	Origin of a transferred thing	I flew in from **Pittsburgh**.
Goal	Destination of a transferred thing	Go to **hell**!

Verb Subcategorization

Verbs have sets of allowed args. Could have many sets of VP rules. Instead, have a SUBCAT feature, marking sets of allowed arguments:

- +none -- Jack laughed
- +np -- Jack found a key
- +np+np -- Jack gave Sue the paper
- +vp:inf -- Jack wants to fly
- +np+vp:inf -- Jack told the man to go
- +vp:ing -- Jack keeps hoping for the best
- +np+vp:ing -- Jack caught Sam looking at his desk
- +np+vp:base -- Jack watched Sam look at his desk
- +np+pp:to -- Jack gave the key to the man
- +pp:loc -- Jack is at the store
- +np+pp:loc -- Jack put the box in the corner
- +pp:mot -- Jack went to the store
- +np+pp:mot -- Jack took the hat to the party
- +adjp -- Jack is happy
- +np+adjp -- Jack kept the dinner hot
- +sthat -- Jack believed that the world was flat
- +sfor -- Jack hoped for the man to win a prize

50-100 possible ***frames*** for English; a single verb can have several.

(Notation from James Allen "Natural Language Understanding")

Thematic Grid or Case Frame

- Example: break
  - The child broke the vase. < agent theme >

    subj obj
  - The child broke the vase with a hammer.

    < agent theme instr >

    subj obj PP
  - The hammer broke the vase. < theme instr >

    obj subj
  - The vase broke. < theme >

    subj

Thematic Grid or Case Frame

- Example: break
  - The child broke the vase. < agent theme >
    subj obj
  - The child broke the vase with a hammer.
    < agent theme instr >
    subj obj PP
  - The hammer broke the vase. < theme instr >
    obj subj
  - The vase broke. < theme >
    subj

The Thematic Grid or Case Frame shows

- How many arguments the verb has
- What roles the arguments have
- Where to find each argument
  - For example, you can find the agent in the subject position

Diathesis Alternation:

a change in the number of arguments or the grammatical relations associated with each argument

Chris gave a book to Dana.	<	agent subj	theme obj	goal PP	>
A book was given to Dana by Chris.	<	agent PP	theme subj	goal PP	>
Chris gave Dana a book.	<	agent subj	theme obj2	goal obj	>
Dana was given a book by Chris.	<	agent PP	theme obj	goal subj	>

Speech Acts

- ***Mood*** of a sentence indicates relation between speaker and the concept (proposition) defined by the LF
- There can be operators that represent these relations:
  - ASSERT: the proposition is proposed as a fact
  - YN-QUERY: the truth of the proposition is queried
  - COMMAND: the proposition describes a requested action
  - WH-QUERY: the proposition describes an object to be identified

ASSERT (Declarative mood)

- *The man ate a peach*

  ASSERT(The x | Man(x) . (A y | Peach(y) . Eat(x,y)))

YN-QUERY (Interrogative mood)

- *Did the man eat a peach?*

  YN-QUERY(The x | Man(x) . (A y | Peach(y) . Eat(x,y)))

COMMAND (Imperative mood)

- *Eat a peach, (man).*

  COMMAND(A y | Peach(y) . Eat(*HEARER*,y))

WH-QUERY

- *What did the man eat?*

  WH-QUERY(The x | Man(x) . (WH y | Thing(y) . Eat(x,y)))
- One of a whole set of new quantifiers for wh-questions:
  - *What*: WH x | Thing(x)
  - *Which dog*: WH x | Dog(x)
  - *Who*: WH x | Person(x)
  - *How many men*: HOW-MANY x | Man(x)

Embedded Sentences

- *The man who ate a peach left*

```
(ASSERT
 (LEAVE
   [AGENT
    (THE m1 |
       MAN(m1) .
            (EAT
             [AGENT m1]
             [THEME (A p1 | PEACH(p1))]))]))
```

Other complications

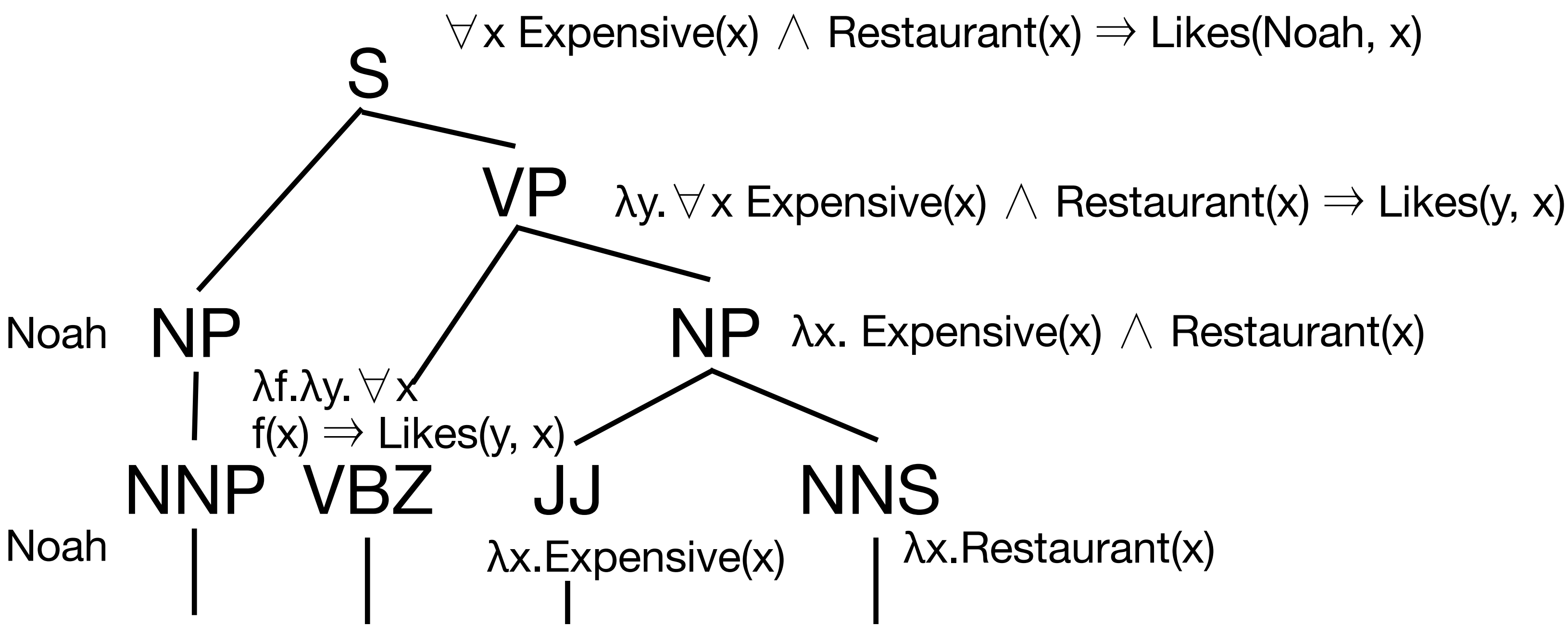
- Modal verbs: non-transparency for truth of subordinate clause: *Sue thinks that John loves Sandy*
- Tense/Aspect
- Plurality
- Etc.

- *You can take this too far…*

Analyzing NL into meaning

- First, syntactic analysis.
- Then, assign meaning in syntax-directed fashion.

Connecting FOPC to Syntax



- *Noah likes expensive restaurants.*
- $\forall x\ Restaurant(x) \wedge Expensive(x) \Rightarrow Likes(Noah, x)$

Analyzing NL into meaning

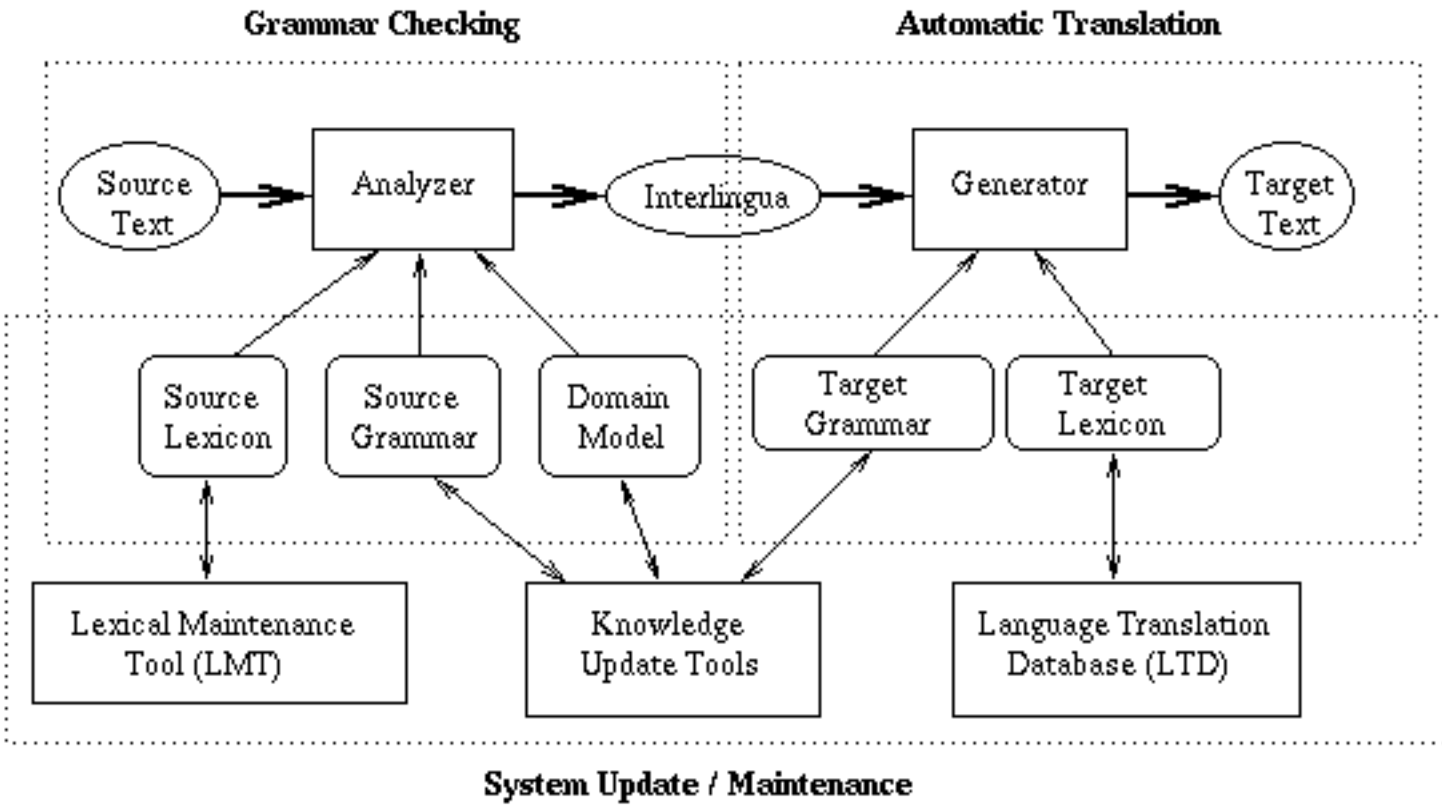
- First, syntactic analysis.
- Then, assign meaning in syntax-directed fashion.
  - Interleaving generally a very good idea

- For MT, don't need to worry about *grounding*. You *would* if you were talking to a robot.
- Can also ignore many discourse issues. Eg, assume pronouns just translate as pronouns.

CMU KANT system

- Produced Catalyst system for Caterpillar
  - Bulldozer manuals in N languages
- Controlled input language
  - Checker/disambiguator, incl domain semantics
- Tomita parser
- LFG-like grammar, pseudo-unification
- Achieved human level translation!
  - (Many people don't realize there ever was a successful KBMT system)

KANTOO system diagram



The KANT Interlingua

- Explicit word senses represented as single terms
- No generalized quantifiers (represented as features)
- Otherwise, very similar to the LF event notation with semantic roles
- (Demonstration)

One of the most successful of these institutions is BancoSol in Bolivia.

```
(*A-BE
   (FORM FINITE)
   (TENSE PRESENT)
   (MOOD DECLARATIVE)
   (PUNCTUATION PERIOD)
   (IMPERSONAL -)
   (THEME
      (*G-PARTITIVE
         (SUBSTANCE
            (*G-PARTITIVE
               (SUBSTANCE
                  (*O-INSTITUTION
                     (UNIT -)
                     (NUMBER PLURAL)
                     (REFERENCE DEFINITE)
                     (DISTANCE NEAR)
                     (PERSON THIRD)))
               (ADJECTIVE
                  (*P-SUCCESSFUL
                     (DEGREE SUPERLATIVE)))))
         (QUANTIFIER (*QUANT-ONE))))
   (PREDICATE
      (*PROP-BANCOSOL
         (NUMBER SINGULAR)
         (IMPLIED-REFERENCE +)
         (PERSON THIRD)
         (UNIT -)
         (Q-MODIFIER
            (*K-IN
               (OBJECT
                  (*PROP-BOLIVIA
                     (UNIT -)
                     (NUMBER SINGULAR)
                     (IMPLIED-REFERENCE +)
                     (PERSON THIRD))))))))
```

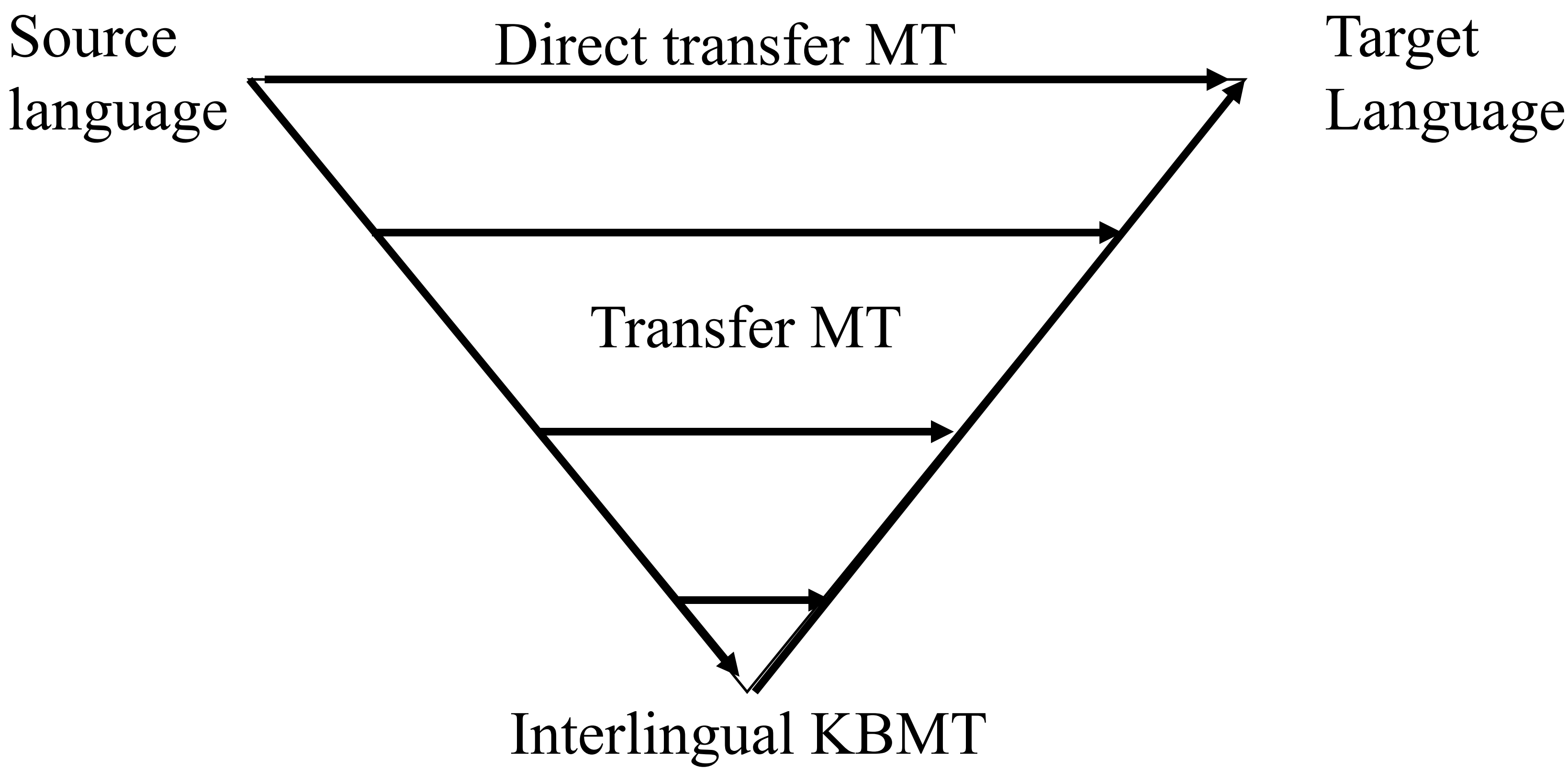
ACCION International is a U.S.-based private non-profit organization that currently provides technical assistance to a network of institutions in thirteen countries in Latin America and six cities in the United States.

```
(*A-BE
   (FORM FINITE)
   (TENSE PRESENT)
   (MOOD DECLARATIVE)
   (PUNCTUATION PERIOD)
   (IMPERSONAL -)
   (THEME
      (*PN-ACCION-INTERNATIONAL
         (NUMBER SINGULAR)
         (IMPLIED-REFERENCE +)
         (PERSON THIRD)
         (UNIT -)))
   (PREDICATE
      (*O-ORGANIZATION
         (UNIT -)
         (NUMBER SINGULAR)
         (REFERENCE INDEFINITE)
         (PERSON THIRD)
         (REL-QUAL
            (*G-QUALIFYING-EVENT
               (EVENT
                  (*A-PROVIDE
                     (MISSING-VERB -)
                     (FORM FINITE)
                     (TENSE PRESENT)
                     (MOOD DECLARATIVE)
                     (IMPERSONAL -)
                     (ARGUMENT-CLASS AGENT+THEME+RECIPIENT)
                     (MANNER
                        (*M-CURRENTLY
                           (POSITION PREVERBAL)
                           (UNIT -)
                           (DEGREE POSITIVE)))
                     (RECIPIENT
                        (*O-NETWORK
                           (UNIT -)
                           (NUMBER SINGULAR)
                           (REFERENCE INDEFINITE)
                           (PERSON THIRD)
                           (Q-MODIFIER
                              (*K-OF
                                 (OBJECT
```

```
      (DEGREE POSITIVE))))
(RECIPIENT
      (*O-NETWORK
         (UNIT -)
         (NUMBER SINGULAR)
         (REFERENCE INDEFINITE)
         (PERSON THIRD)
         (Q-MODIFIER
            (*K-OF
               (OBJECT
                  (*O-INSTITUTION
                     (UNIT -)
                     (NUMBER PLURAL)
                     (REFERENCE NO-REFERENCE)
                     (PERSON THIRD)
                     (Q-MODIFIER
                        (*K-IN
                           (OBJECT
                              (*G-COORDINATION
                                 (UNIT -)
                                 (PERSON THIRD)
                                 (NUMBER PLURAL)
                                 (REFERENCE NO-REFERENCE)
                                 (CONJUNCTION (*CONJ-AND))
                                 (CONJUNCTS
                                    (:MULTIPLE
                                       (*O-COUNTRY
                                          (NUMBER PLURAL)
                                          (UNIT -)
                                          (PERSON THIRD)
                                          (IMPLIED-REFERENCE +)
                                          (Q-MODIFIER
                                             (*K-IN
                                                (OBJECT
                                                   (*PROP-LATIN-AMERICA
                                                      (UNIT -)
                                                      (NUMBER SINGULAR)
                                                      (IMPLIED-REFERENCE +)
                                                      (PERSON THIRD)))))
                                          (QUANTITY
                                             (*C-DECIMAL-NUMBER
                                                (INTEGER "13")
                                                (NUMBER-FORM ALPHABETIC)
                                                (NUMBER-TYPE CARDINAL))))
                                       (*O-CITY
```

```
                                                              (QUANTITY
                                                                 (*C-DECIMAL-NUMBER
                                                                    (INTEGER "13")
                                                                    (NUMBER-FORM ALPHABETIC)
                                                                    (NUMBER-TYPE CARDINAL))))
                                                           (*O-CITY
                                                              (NUMBER PLURAL)
                                                              (UNIT -)
                                                              (PERSON THIRD)
                                                              (IMPLIED-REFERENCE +)
                                                              (Q-MODIFIER
                                                                 (*K-IN
                                                                    (OBJECT
                                                                       (*PROP-UNITED-STATES
                                                                          (UNIT -)
                                                                          (NUMBER SINGULAR)
                                                                          (REFERENCE DEFINITE)
                                                                          (PERSON THIRD)))))
                                                              (QUANTITY
                                                                 (*C-DECIMAL-NUMBER
                                                                    (INTEGER "6")
                                                                    (NUMBER-FORM ALPHABETIC)
                                                                    (NUMBER-TYPE CARDINAL))))))))))))))
                    (THEME
                       (*O-ASSISTANCE
                          (UNIT -)
                          (NUMBER SINGULAR)
                          (REFERENCE NO-REFERENCE)
                          (PERSON THIRD)
                          (ATTRIBUTE
                             (*P-TECHNICAL
                                (DEGREE POSITIVE)))))
                    (AGENT
                       (*G-GAPPED-ARGUMENT
                          (GAPPED +)))))
              (EXTENT (*REL-THAT))))
        (ATTRIBUTE
           (*G-COORDINATION
              (CONJUNCTION NULL)
              (CONJUNCTS
                 (:MULTIPLE
                    (*P-US-BASED (DEGREE
                      POSITIVE))
                    (*P-PRIVATE (DEGREE
                      POSITIVE))
                    (*P-NON-PROFIT (DEGREE
                      POSITIVE)))))))))
```

Types of MT technologies



[After Vauquois]

Generating NL from meaning

- Not trivial, but not as hard as parsing/ interpretation (if meaning representation well-designed)
- MT can again mostly avoid some major issues
  - Content selection
  - Discourse coherence

Generating from meaning

- Need to express content while obeying linguistic constraints
- A form of planning, vs. analysis
  - Backtracking may be necessary, if linguistic constraints become unsatisfiable

NLG for KBMT

- Template-based generators
  - Weather reports?
- CGI LanguageCraft generator
  – Case-frame based representation
- CMU KANT GenKit generator
  – LFG-like syntax, frame-style semantics
- ISI Pangloss Penman generator
  – Systemic grammar, planner in LISP

Using Case frames for NLG:

- Example: break
  - The child broke the vase. < agent theme >
    subj obj
- The child broke the vase with a hammer.
  < agent theme instr >
  subj obj PP
- The hammer broke the vase. < theme instr >
  obj subj
- The vase broke. < theme >
  subj

The Thematic Grid or Case Frame shows

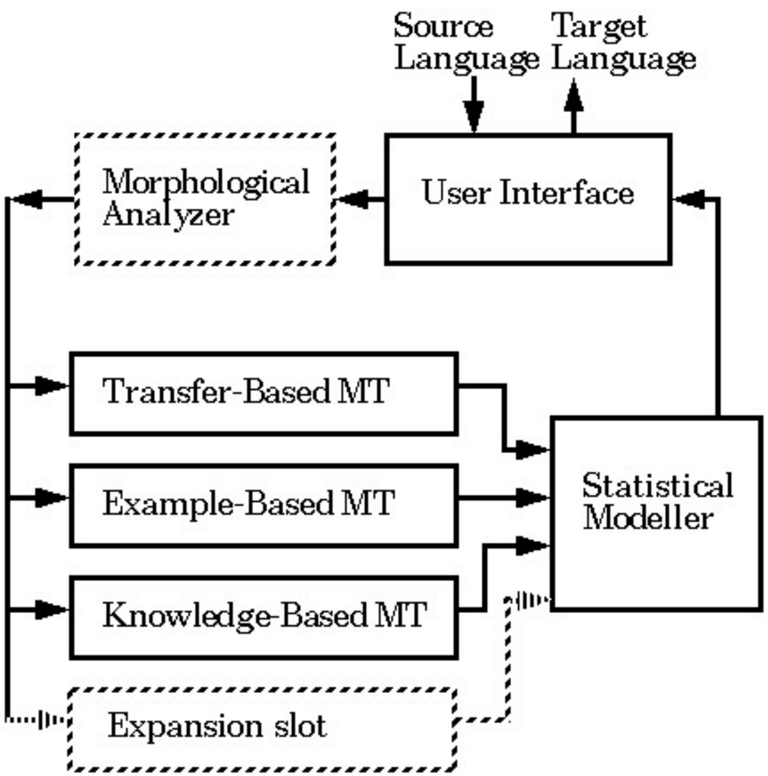
- How many arguments the verb has
- What roles the arguments have
- Where to find each argument
  - For example, you can find the agent in the subject position

Issues with KBMT

- Only really possible in limited domains
  - But not necessarily trivial ones
- Knowledge engineering is very expensive

- Interlingua: not clear that a universal interlingua is actually possible
  - But it doesn't really have to be universal in practice

Multi-Engine MT



- Apply several MT engines to each input; use statistical language modeller to select best combination of outputs.
- Goal is to combine strengths, and avoid weaknesses.
- Along all dimensions: domain limits, quality, development time/cost, run-time speed, etc.
- Used in Diplomat, Tongues, LingWear, Nespole, NICE, etc.

Example MEMT "chart"

El punto de descarge **The drop-off point**	*se cumplirá en* will comply with	*el puente Agua Fria* The cold Bridgewater
El punto de descarge The discharge point	*se cumplirá en* will self comply in	*el puente Agua Fria* **the "Agua Fria" bridge**
El punto de descarge Unload of the point	*se cumplirá en* **will take place at**	*el puente Agua Fria* the cold water of bridge

Current RBMT/KBMT

- Still used in industry, especially where high-precision domain-specific MT is needed
- No research funding

- But note that “statistical” MT systems often include rule-based components, esp. morphology

Questions?