
SWE1017

Natural Language Processing

Venue: AB2-204

Topic: Machine Translation

***Prof. Tulasi Prasad Sariki,
SCSE, VIT Chennai Campus***

www.learnersdesk.weebly.com



Contents

- ❑ History of Machine Translation
- ❑ Machine Translation: Where are we today?
- ❑ How Does MT Work?
- ❑ Core Challenges of MT
- ❑ Rule-based vs. Data-driven Approaches to MT
- ❑ Statistical MT (SMT)
- ❑ Major Sources of Translation Problems
- ❑ Speech to speech translation

History of Machine Translation

1946: MT is one of the 1st conceived applications of modern computers (Alan Turing)

1954: The “Georgetown Experiment” demonstrations of Russian-English MT

Late 1950s and early 1960s: MT fails to scale up to “real” systems

1966: ALPAC Report: MT recognized as an extremely difficult, “AI-complete” problem. Funding disappears

1968: SYSTRAN founded

1985: CMU “Center for Machine Translation” (CMT) founded

Late 1980s and early 1990s: Field dominated by rule-based approaches – KBMT, KANT, Eurotra, etc.

History of Machine Translation

1992: “Noisy Channel” Statistical MT models invented by IBM (CANDIDE)

Mid 1990s: First major DARPA MT Program. PANGLOSS

Late 1990s: Major Speech-to-Speech MT demonstrations: C-STAR

1999: JHU Summer Workshop results in GIZA

2000s: Large DARPA Funding Programs – TIDES and GALE

2003: Och et al introduce Phrase-based SMT. PHARAOH

2006: Google Translate is launched

2007: Koehn et al release MOSES

2008: a text/SMS translation service for mobiles in Japan

2009: mobile phone with built-in speech-to-speech translation facility for English and Japanese

2012: Google announced that Google Translate

MT: Where are we today?

Age of Internet & Globalization – great demand for translation services and MT

- ❑ Multiple official languages of UN, EU, Canada, etc.
- ❑ Commercial demand from increasing number of global enterprises
 - ❑ (Microsoft, IBM, Intel, Apple, E-bay, Amazon, etc.)
- ❑ Language and translation services business sector estimated at \$15 Billion worldwide in 2008 and growing at a healthy pace

Economic incentive and demand is still focused primarily within G-8 languages, but growing in emerging markets (BRIC: Brazil, Russia, India, China), Arabic, and more...

MT: Where are we today?

Some fairly decent commercial products in the market for these language pairs

- ❑ Primarily a product of rule-based systems after many years of development
- ❑ New generation of data-driven “statistical” MT: Google, Microsoft, Language Weaver

Web-based (mostly free) MT services: Google, Babelfish, others...

Pervasive MT between many language pairs still non-existent, but Google is trying to change that!

How Does MT Work?

All modern MT approaches are based on building translations for complete sentences by putting together smaller pieces of translation

Core Questions:

- ❑ What are these smaller pieces of translation?
- ❑ Where do they come from?
- ❑ How does MT put these pieces together?
- ❑ How

Core Challenges of MT

Ambiguity and Language Divergences:

- ❑ Human languages are highly ambiguous, and differently in different languages
- ❑ Ambiguity at all “levels”: lexical, syntactic, semantic, language-specific constructions and idioms

Amount of required knowledge:

- ❑ Translation equivalencies for vast vocabularies
- ❑ Syntactic knowledge (how to map syntax of one language to another), plus more complex language divergences (semantic differences, constructions and idioms, etc.)
- ❑ **How**

Rule-based vs. Data-driven Approaches to MT

What are the pieces of translation? Where do they come from?

- ❑ **Rule-based:** large-scale “clean” word translation lexicons, manually constructed over time by experts
- ❑ **Data-driven:** broad-coverage word and multi-word translation lexicons, learned automatically from available sentence-parallel corpora

How does MT put these pieces together?

- ❑ **Rule-based:** large collections of rules, manually developed over time by human experts, that map structures from the source to the target language
- ❑ **Data-driven:**

Rule-based vs. Data-driven Approaches to MT

How does the MT system pick the correct (or best) translation among many options?

- ❑ **Rule-based:** Human experts encode preferences among the rules designed to prefer creation of better translations
- ❑ **Data-driven:**

Rule-based vs. Data-driven Approaches to MT

Why have the data-driven approaches become so popular?

- ❑ Increasing amounts of sentence-parallel data are constantly being created on the web
- ❑ Advances in machine learning algorithms
- ❑ Computational power of today's computers can train systems on these massive amounts of data and can perform these massive search-based translation computations when translating new texts
- ❑ Building and maintaining rule-based systems is too difficult, expensive and time-consuming
- ❑ In

Statistical MT (SMT)

Data-driven, most dominant approach in current MT research

Proposed by IBM in early 1990s: a direct, purely statistical, model for MT

Evolved from word-level translation to phrase-based translation

Main Ideas:

- ❑ **Training:** statistical “models” of word and phrase translation equivalence are learned automatically from bilingual parallel sentences, creating a bilingual “database” of translations
- ❑ **Decoding:** new sentences are translated by a program (the decoder), which matches the source words and phrases with the database of translations, and searches the “space” of all possible translation combinations.

Statistical MT (SMT)

Main steps in training phrase-based statistical MT:

- ❑ Create a sentence-aligned parallel corpus
- ❑ **Word Alignment:** train word-level alignment models (GIZA++)
- ❑ **Phrase Extraction:** extract phrase-to-phrase translation correspondences using heuristics (Moses)
- ❑ **Minimum Error Rate Training (MERT):** optimize translation system parameters on development data to achieve best translation performance

Attractive: completely automatic, no manual rules, much reduced manual labor

Statistical MT (SMT)

Main drawbacks:

- ❑ Translation accuracy levels vary widely
- ❑ Effective only with large volumes (several mega-words) of parallel text
- ❑ Broad domain, but domain-sensitive
- ❑ Viable only for limited number of language pairs!

Impressive progress in last 5-10 years!

Statistical MT: Major Challenges

Current approaches are too naïve and “direct”:

- ❑ Good at learning word-to-word and phrase-to-phrase correspondences from data
- ❑ Not good enough at learning how to combine these pieces and reorder them properly during translation
- ❑ Learning general rules requires much more complicated algorithms and computer processing of the data
- ❑ The space of translations that is “searched” often doesn’t contain a perfect translation
- ❑ The fitness scores that are used aren’t good enough to always assign better scores to the better translations ✉ we don’t always find the best translation even when it’s there!
- ❑ MERT is brittle, problematic and metric-dependent!

Statistical MT: Major Challenges

Solutions:

- ❑ Google solution: more and more data!
- ❑ Research solution: “smarter” algorithms and learning methods

Rule-based vs. Data-driven MT

We thank all participants of the whole world for their comical and creative drawings; to choose the victors was not easy task!

Click here to see work of winning European of these two months, and use it to look at what the winning of USA sent us.

Rule-based

We thank all the participants from around the world for their designs cocasses and creative; selecting winners was not easy!

Click here to see the artwork of winners European of these two months, and disclosure to look at what the winners of the US have been sending.

Data-driven

Major Sources of Translation Problems

Lexical Differences:

- ❑ Multiple possible translations for SL word, or difficulties expressing SL word meaning in a single TL word

Structural Differences:

- ❑ Syntax of SL is different than syntax of the TL: word order, sentence and constituent structure

Differences in Mappings of Syntax to Semantics:

- ❑ Meaning in TL is conveyed using a different syntactic structure than in the SL

Idioms

How to Tackle the Core Challenges

Manual Labor: 1000s of person-years of human experts developing large word and phrase translation lexicons and translation rules.

Example: Systran's RBMT systems.

Lots of Parallel Data: data-driven approaches for finding word and phrase correspondences automatically from large amounts of sentence-aligned parallel texts. Example: Statistical MT systems.

Learning Approaches: learn translation rules automatically from small amounts of human translated and word-aligned data. Example: AVENUE's Statistical XFER approach.

Simplify the Problem: build systems that are limited-domain or constrained in other ways. Examples: CATALYST, NESPOLE!.

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 (Example-Based) MT
- ❑ GP HQ: Human-in-the-loop (Post-editing)

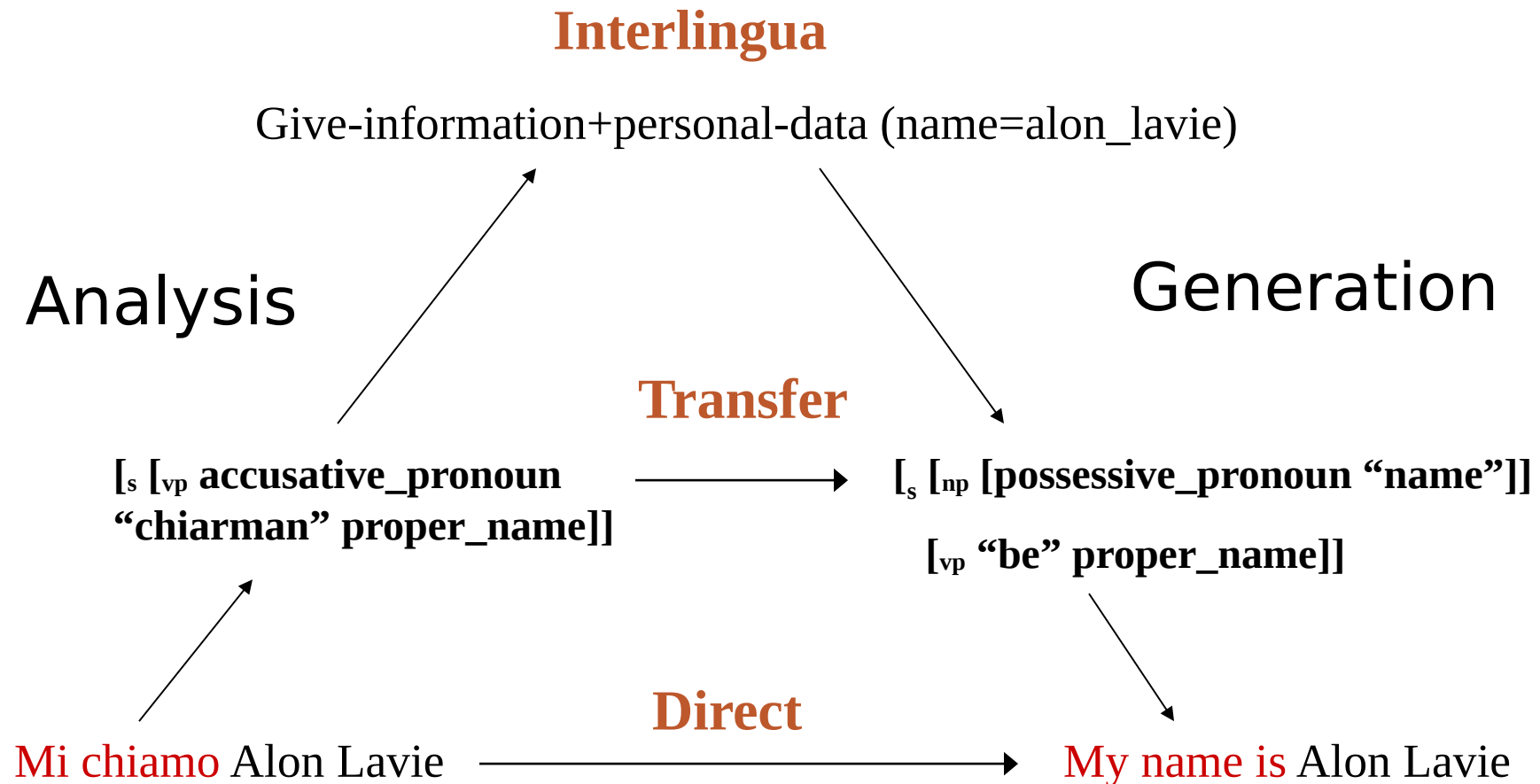
Types of MT Applications:

Assimilation: multiple source languages, uncontrolled style/topic. General purpose MT, no customization. (Google Translate)

Dissemination: one source language, controlled style, single topic/domain. Customized RBMT or SMT. (Safaba)

Communication: Lower quality may be okay, but system robustness, real-time required. (Jibiggo)

Approaches to MT: Vaquois MT Triangle



Direct Approaches

No intermediate stage in the translation

First MT systems developed in the 1950's-60's (assembly code programs)

- ❑ Morphology, bi-lingual dictionary lookup, local reordering rules
- ❑ “Word-for-word, with some local word-order adjustments”

Modern Approaches:

- ❑ Phrase-based Statistical MT (SMT)
- ❑ Example-based MT (EBMT)

EBMT Paradigm

New Sentence (Source): Yesterday, 200 delegates met with President Clinton.

Matches to Source Found

Yesterday, 200 delegates met behind closed doors...

Gestern trafen sich 200 Abgeordnete hinter verschlossenen...

Difficulties with President Clinton...

Schwierigkeiten mit Praesident Clinton...

Alignment (Sub-sentential)

Yesterday, 200 delegates met behind closed doors...

Gestern trafen sich 200 Abgeordnete hinter verschlossenen...

Difficulties with President Clinton over...

Schwierigkeiten mit Praesident Clinton...

Translated Sentence (Target): Gestern trafen sich 200 Abgeordnete mit Praesident Clinton.

Analysis and Generation Main Steps

Analysis:

- Morphological analysis (word-level) and POS tagging
- Syntactic analysis and disambiguation (produce syntactic parse-tree)
- Semantic analysis and disambiguation (produce symbolic frames or logical form representation)
- Map to language-independent Interlingua

Generation:

- Generate semantic representation in TL
- Sentence Planning: generate syntactic structure and lexical selections for concepts
- Surface-form realization: generate correct forms of words

Transfer Approaches

Syntactic Transfer:


- ❑ Analyze SL input sentence to its syntactic structure (parse tree)
- ❑ Transfer SL parse-tree to TL parse-tree (various formalisms for mappings)
- ❑ Generate TL sentence from the TL parse-tree

Semantic Transfer:



- ❑ Analyze SL input to a language-specific semantic representation (i.e., Case Frames, Logical Form)
- ❑ Transfer SL semantic representation to TL semantic representation
- ❑ Generate syntactic structure and then surface sentence in the TL

Transfer Approaches (Pros & Cons)

Syntactic Transfer:

- No need for semantic analysis and generation
- Syntactic structures are general, not domain specific  Less domain dependent, can handle open domains
- Requires word translation lexicon

Semantic Transfer:

- Requires deeper analysis and generation, symbolic representation of concepts and predicates  difficult to construct for open or unlimited domains
- Can better handle non-compositional meaning structures  can be more accurate
- No word translation lexicon – generate in TL from symbolic concepts

Knowledge-based Interlingual MT

The classic “deep” Artificial Intelligence approach:

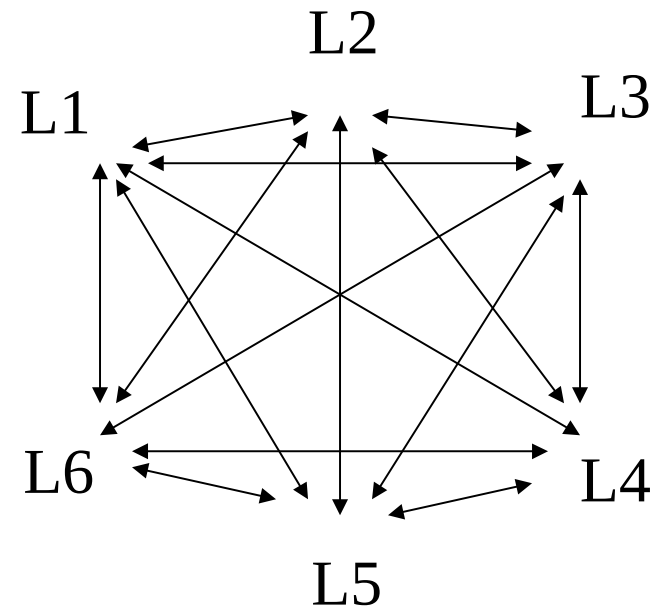
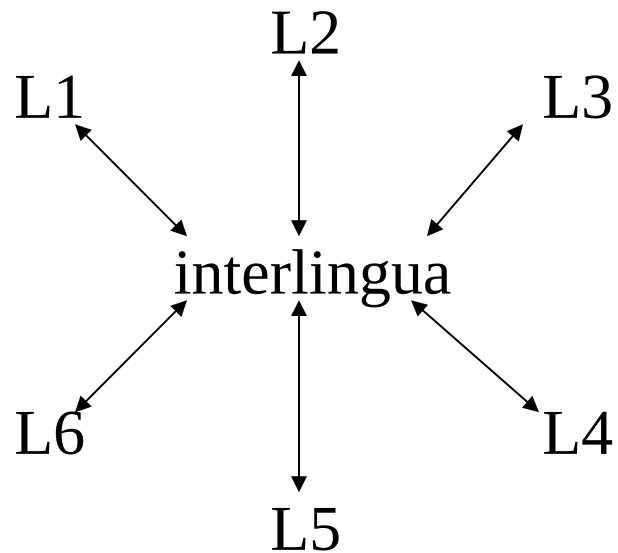
- ❑ Analyze the source language into a detailed symbolic representation of its meaning
- ❑ Generate this meaning in the target language

“Interlingua”: one single meaning representation for all languages

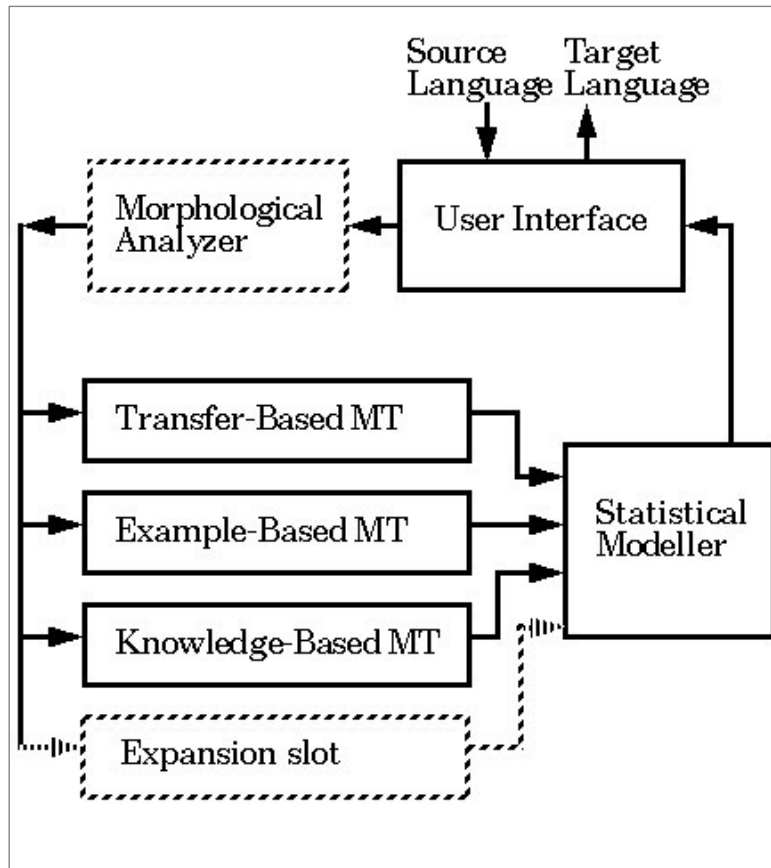
- ❑ Nice in theory, but extremely difficult in practice:
 - ❑ What kind of representation?
 - ❑ What is the appropriate level of detail to represent?
 - ❑ How to ensure that the interlingua is in fact universal?

Interlingua versus Transfer

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



Multi-Engine MT



Apply several MT engines to each input in parallel

Create a combined translation from the individual translations

Goal is to combine strengths, and avoid weaknesses.

Along all dimensions: domain limits, quality, development time/cost, run-time speed, etc.

Various approaches to the problem

Speech-to-Speech MT

Speech just makes MT (much) more difficult:

- ❑ Spoken language is messier
 - ❑ False starts, filled pauses, repetitions, out-of-vocabulary words
 - ❑ Lack of punctuation and explicit sentence boundaries
- ❑ Current Speech technology is far from perfect

Need for speech recognition and synthesis in foreign languages

Robustness: MT quality degradation should be proportional to SR quality

Tight Integration: rather than separate sequential tasks, can SR + MT be integrated in ways that improves end-to-end performance?

END