

# Machine Translation Overview

Alon Lavie  
Language Technologies Institute  
Carnegie Mellon University

LTI Immigration Course  
August 22, 2011

# Machine Translation: History

- **1946:** MT is one of the first conceived applications of modern computers (A.D. Booth, Alan Turing)
- **1954:** The “Georgetown Experiment” Promising “toy” 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.
- **1992:** “Noisy Channel” Statistical MT models invented by IBM researchers (Brown, Della Pietra, et al.). 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

# Machine Translation: Where are we today?

- Age of Internet and Globalization – great demand for translation services and MT:
  - Multiple official languages of UN, EU, Canada, etc.
  - Software Localization and documentation dissemination for large manufacturers (Microsoft, Intel, Apple, EBay, ALCOA, etc.)
  - Language and translation services business sector estimated at \$26 Billion worldwide in 2010 and growing at a healthy pace
  - Volume of online content growing exponentially
- Economic incentive is still primarily within a small number of language pairs
- Some fairly decent commercial products in the market for these language pairs
  - Product of rule-based systems after many years of development: SYSTRAN, PROMT, others...
  - New generation of data-driven “statistical” MT systems: SDL/Language Weaver, Asia Online, others...
- Web-based (mostly free) MT services: Google, MS-Bing, Babelfish, others...
- Pervasive MT between many language pairs still non-existent, but some significant progress in recent years

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

# 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 (several 100k words and phrases)
  - Syntactic knowledge (how to map syntax of one language to another), plus more complex language divergences (semantic differences, constructions and idioms, etc.)
  - **How** do you acquire and construct a knowledge base that big that is (even mostly) correct and consistent?

# 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!

# 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**: a computer algorithm that explores millions of possible ways of putting the small pieces together, looking for the translation that statistically looks best

# 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:** a variety of fitness and preference scores, many of which can be learned from available training data, are used to model a total score for each of the millions of possible translation candidates; algorithm then selects and outputs the best scoring translation



# Rule-based vs. Data-driven Approaches to MT

- Why have the data-driven approaches become so popular?
  - We can now do this!
    - 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 many scenarios, it actually works better!

# Statistical MT (SMT)

- Data-driven, most dominant approach in current MT research
- Originally 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: 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!
- **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.

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.

Rule-based

Data-driven

# Representative Example: Google Translate

- <http://translate.google.com>

# Google Translate

[Web](#) [Images](#) [Maps](#) [News](#) [Video](#) [Gmail](#) [more](#) ▼ [Help](#)

**Google**  
Translate BETA

[Home](#) [Text and Web](#) [Translated Search](#) [Tools](#)

---

### Translate text or webpage

Enter text or a webpage URL.

El TPIY pone en libertad al ex presidente serbio Milutinovic

El ex mandatario afrontaba cargos por crímenes contra la humanidad durante la guerra de Kosovo.- Otros cinco altos cargos serbios, condenados a entre 15 y

Translation: Spanish » English

The ICTY set free the former Serbian President Milutinovic

The former president was facing charges for crimes against humanity during the Kosovo war .- Five other senior Serbs, sentenced to between 15 and 22 years in prison.

Spanish > English [swap](#)

[+ Suggest a better translation](#)

---

[Google Home](#) - [About Google Translate](#)

©2009 Google

# Google Translate



The screenshot shows the Google Translate web interface. At the top, there are navigation links for Web, Images, Video, Maps, News, Shopping, Gmail, and more. The main heading is "Google translate" with tabs for Home, Text and Web, Translated Search, and Tools. The page title is "Translate text or webpage". Below this, there is a text input field containing Chinese text about reducing advertising costs during a financial crisis. To the right, the translated English text is displayed. The browser's address bar is visible at the bottom, showing the URL "http://www.cninfo.net".

Web Images Video Maps News Shopping Gmail more ▾ Help ▲

Google translate Home Text and Web Translated Search Tools

**Translate text or webpage**

Enter text or a webpage URL

金融危机下如何降低广告成本

经济危机席卷全球，如何节约成本，度过经济寒冬？网上推广成为中小企业的的首选，通过电子商务，节约成本、吸引买家，成为中国企业的重点选择。中国行业信息网（[www.cninfo.net](http://www.cninfo.net)）百万网络会员目前更加活跃，就充分说明了这一点。

如果您还不是中国行业信息网（[www.cninfo.net](http://www.cninfo.net)）的会员，请立即登录注册：

注册页面：<http://www.cninfo.net/company/reg.aspx>

只需三分钟，您就可以成功注册，您可以建立自己的网上商城，加入合作商友，发布供求信息、人才信息，阅读最新经济新闻，也可以在论坛交流……总之，中国行业信息的大平台，旨在将千万企业聚集起来，让交易更简单、让生意更好做，让大家共同抱团取暖，度过经济寒冬，迎来财富的春天！

现在登录中国行业信息网：[www.cninfo.net](http://www.cninfo.net)，并加入收藏夹。

这样您就可以开始您的网络历程了，从资讯指导、到电子商务、网上推广……一站搞掂！

让中国行业信息网给您带财富的春天！

中国行业信息网客服部

Translation: Chinese » English

How the financial crisis to reduce advertising costs

The economic crisis sweeping the globe, how the cost savings, through the economic winter? Online promotion to become the first choice for small and medium enterprises, through e-commerce, reduce costs and attract buyers to become an important choice for Chinese enterprises. China Industry Information Net ([www.cninfo.net](http://www.cninfo.net)) network of millions of active members at present more fully illustrates this point.

If you're not a trade information network in China ([www.cninfo.net](http://www.cninfo.net)) member, please log on for registration:

Registration page:  
<http://www.cninfo.net/company/reg.aspx>

Takes only three minutes, you can successfully registered, you can create your own online mall, to join the Friends of partners, the supply and demand release of information, human resources

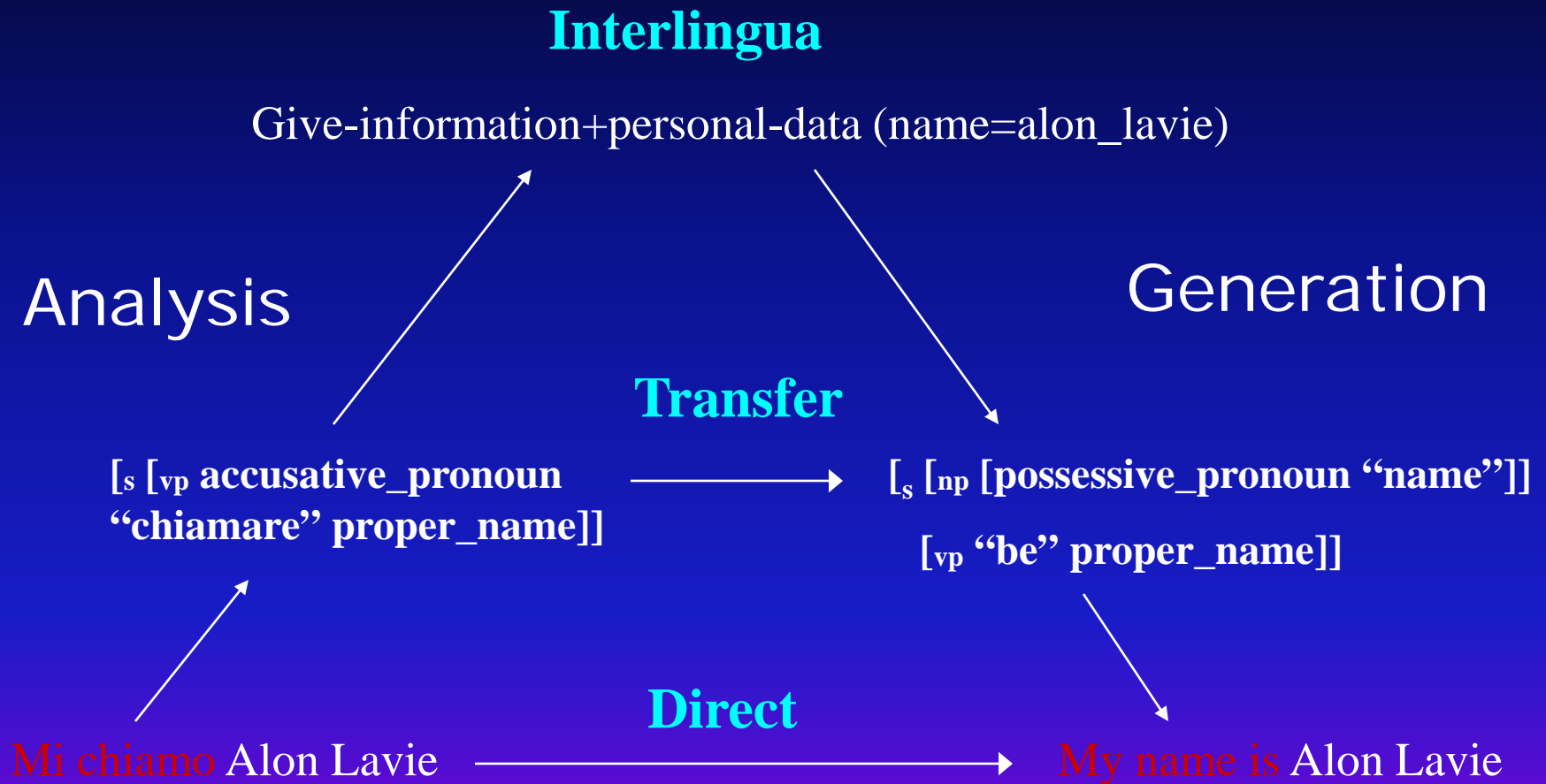
<http://www.cninfo.net>

# Types of MT Applications:

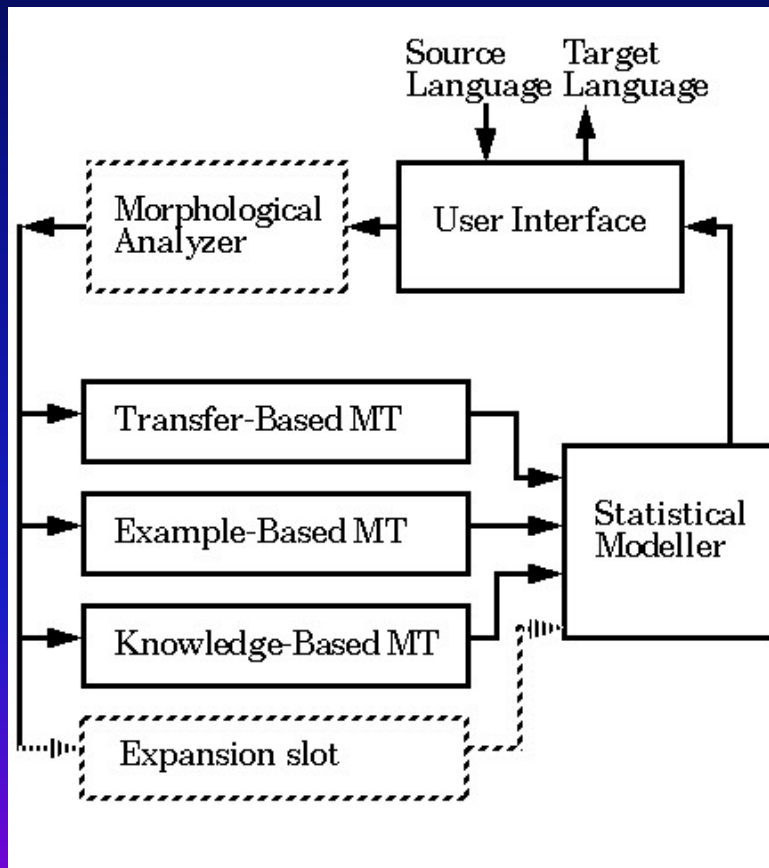
- **Assimilation:** multiple source languages, uncontrolled style/topic. General purpose MT, no semantic analysis. (GP FA or GP HQ)
- **Dissemination:** one source language, controlled style, single topic/domain. Special purpose MT, full semantic analysis. (FA HQ)
- **Communication:** Lower quality may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



# Approaches to MT: Vaquois MT Triangle



# 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?

# MT at the LTI

- LTI originated as the Center for Machine Translation (CMT) in 1985
- MT continues to be a prominent sub-discipline of research with the LTI
- Active research on all main approaches to MT
- Leader in the area of speech-to-speech MT
- Multi-Engine MT (MEMT)
- MT Evaluation (METEOR)
- Spin-off Companies:
  - Jibbigio (speech translation on mobile devices)
  - Safaba (MT solutions for enterprises and LSPs)

# MT Faculty at LTI

- Alon Lavie
- Stephan Vogel
- Ralf Brown
- Jaime Carbonell
- Lori Levin
- Noah Smith
- Alan Black
- Florian Metze
- Alex Waibel
- Teruko Mitamura
- Eric Nyberg

# Summary

- Main challenges for current state-of-the-art MT approaches - Coverage and Accuracy:
  - Acquiring broad-coverage high-accuracy translation lexicons (for words and phrases)
  - learning structural mappings between languages from parallel word-aligned data
  - overcoming syntax-to-semantics differences and dealing with constructions
  - Stronger Target Language Modeling
  - Context-dependent modeling and adaptation
  - Novel algorithms for model acquisition and decoding

# Questions...