Machine Translation

• **Machine Translation (MT)** systems automatically translate one natural language (the *source* language) to another (the *target* language).

• High-quality MT could greatly enhance global communication.

• But … MT is challenging because it requires full understanding as well as generation.

• Early MT systems ran into such difficulties that MT research stopped for many years! But MT has seen a resurgence.

Early Difficulties for MT

Ich muss nach Hause gehen
*I must go home*

hydraulic ram
*wasser schaf*

out of sight, out of mind
*blind, crazy*

The spirit is willing but the flesh is weak
*The vodka is good but the meat is rotten*

Famous Human Translation Fails

Pepsi slogan “Come alive with the Pepsi Generation” → “Pepsi will bring your ancestors back from the dead”

Kentucky Fried Chicken slogan “finger-lickin’ good” → “eat your fingers off”

Chevy Nova → Nova means “won’t go” in Spanish!

Scandinavian vacuum manufacturer Electrolux used this in an American ad: “Nothing sucks like an Electrolux”

Siemens washing machine: “made in Turkey” → “made by a turkey”

Epic Machine Translation Fail

There’s a serious point here … it’s dangerous to put total trust in an MT system when you have no idea what its output means!
The grieving people of Littleton were burying furnace more victims of the Columbine High School massacres one Monday as investigators said that the two teen gunmen had hoped to kill hundreds of classmates and teachers. Authorities also said year unidentified 18-year-old female being questioned about whether she purchased any firearms used in the rampage is not considered has suspect “at this time”.

The grieving people of Littleton were burying four more victims of the Columbine High School massacre on Monday as investigators said that the two teen gunmen had hoped to kill hundreds of classmates and teachers. Authorities also said an unidentified 18-year-old female being questioned about whether she purchased any firearms used in the rampage is not considered a suspect “at this time”.

The grieving people of Littleton were burying furnace more victims of the Columbine High School massacres one Monday have investigators said that the two teen gunmen had hoped to kill hundreds of classmates and teachers. Authorities also said year unidentified 18-year-old female being questioned but whether she purchased any firearms used in the rampage is not considered has suspect “at this time”.

The grieving people of Littleton were burying four more victims of the Columbine High School massacre on Monday as investigators said that the two teen gunmen had hoped to kill hundreds of classmates and teachers. Authorities also said an unidentified 18-year-old female being questioned about whether she purchased any firearms used in the rampage is not considered a suspect “at this time”.

If you have suggestions, let us know → If you have suggestions, leave knows to us
Another Example

Philadelphia guard Allen Iverson got the 76ers off to a quick start this season. Now he wants to make sure they finish strong. Iverson scored 14 fourth-quarter points on his way to 38 to give Philadelphia a 103-86 victory over the Orlando Magic on Sunday, securing the team's fourth straight win.

Different Types of MT Systems

[Knight, AI Magazine 1997]

Direct Machine Translation

- **Direct machine translation** translates one language directly into another without an intermediate representation.

- Direct MT systems typically use **word for word** translation based on a bilingual dictionary.

- Local word reordering can be used to smooth the target language output.

- Direct MT methods are inherently limited and usually don't perform very well.
Word for Word Translation Problems

Many commercial systems do word for word translation, despite its inherent problems, which include:

- **Lexical Ambiguity**: shot, bat
- **Synonyms & Subtleties**: run, jog, sprint, trot, flee
- **Phrases**: hot dog, real estate, dry run
- **Transposition**: Word order varies a lot, especially between different types of languages.
- **Idioms**: raining cats and dogs, buried the hatchet

Examples (English → French)

He walked across the road.
*Il traversa la rue a pied.*

She drove into town.
*Elle entra dans la ville en voiture.*

They flew from Gatwick.
*Ils partirent par avion de Gatwick.*

Indirect Machine Translation

- **Indirect machine translation** uses an intermediate representation to capture meaning.

  - An *interlingua* is a language-independent meaning representation. The source language is analyzed and its meaning is represented by the interlingua, which is then used to generate output in the target language.

  - The interlingua must be rich enough to capture full understanding!

  - Interlingua-based MT allows for multilingual systems that can translate to and from any pair of languages.

  - Another approach is *transfer* systems that use language-dependent intermediate representations.

Knowledge-based MT (KBMT)

KBMT is knowledge intensive and only feasible for limited domains. But it can perform very well! KBMT requires:

- Bilingual dictionaries with syntactic and semantic information.
- Syntactic grammars.
- A semantic ontology.
- Semantic analyzers.
- Usually an *interlingua* representation.
- Language generation models.
Statistical MT

• The best MT systems today use statistical methods.

• Statistical MT systems are typically trained on parallel text corpora, which consist of source language texts coupled with their manual translations into a target language.

• The Hansard parallel corpus is widely used and contains the proceedings of the Canadian Parliament in both French and English. The European Union also has parallel corpora for many European languages.

• The Web contains many sites that have multiple translations. However, the translations often are not parallel.

Alignment

• In parallel corpora, sentences and paragraphs do not always correspond. For example, one French sentence may be translated as two English sentences.

• Some information (even entire sentences or paragraphs!) may be present in one but missing in the other.

• Sentence alignment algorithms attempt to align the corresponding sentences across parallel texts.

• Word alignment algorithms attempt to align the corresponding words across parallel sentences.

Parallel Text Example

The higher turnover was largely due to an increase in then sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.

La progression des chiffres d'affaires resulte en grande partie de l'accroissement du volume des ventes. L'emploi et le investissements ont egalemt augmentes.

La nouvelle ordonnance federale sur les denrees alimentaires concernant entre autres les eaux minerales, entree en vigueur le ler avril 1988 apres une periode transitoire de deux ans, exige surtout une plus grande constance dans la qualite et une garantie de la purete.

Word Alignment Examples

The letter was sent Tuesday.

Le lettre a ete envoyee le mardi.

Mary did not slap the green witch.

Maria no dio una bofetada a la bruja verde.
Statistical Models for MT

The basic statistical model for MT uses a translation model \( P(S | T) \) and a language model \( P(T) \).

\[
\text{TargetOutput} = \arg\max P(S | T) \times P(T)
\]

The translation model computes the probability that the source text would generate the target text.

The language model computes the probability that the target text would occur in the target language.

Language Generation

Challenges for language generation include:
- lexical choice
- syntactic form (voice, tense, etc.)
- word orderings
- information content

Article Generation:
- I go to school.
- I go to university. (*)
- I saw her on TV.
- I heard him on radio. (*)

Idioms:
- John kicked the bucket.
- The bucket was kicked by John. (*)

Semantic Subtleties:
- President Bush goes fishing and boating in Maine.
- President Bush goes fishing and shipping in Maine. (*)
Using Language Models

Parsers can judge the grammaticality of a sentence, but this is not enough. For example:

John saw Mary on TV.
John viewed Mary in the television. (*)

Statistical language models can be useful for judging both word order and word selection (lexical choice). For example: "en" in Spanish can be translated as both "in" and "on".

I swam in the pool.
I swam on the pool. (*)

Or to judge grammaticality:

I ate bread.
I ate a bread. (*)

Language Generation

The number of possible output sentences can be enormous!

Consider this meaning representation for a Japanese sentence. (Japanese does not mark singular/plural, definiteness, or time.)

ACCUSATION
Agent = she
Theme = THEFT (agent = he  theme = car)

NITROGEN Example

NITROGEN is a language generator that used a dictionary of nouns, verbs, adjectives, and adverbs, plus a hand-built grammar.

NITROGEN found 381,440 English renderings!

Examples:

Her incriminates for him to thieve an automobiles.
There is the accusation of theft of the car by him by her.
She impeaches that he thieve that there was the auto.

Using Language Models to Rank

Statistical methods can be used to rank the choices.

A statistical language model ranked these sentences as the top 5 choices:

She charged that he stole the car.
She charged that he stole the cars.
She charged that he stole cars.
She charged that he stole car.
She charges that he stole the car.
The State of the Art in MT

- Current MT systems are far from perfect, but they are improving and are faster and cheaper than manual translation.

- MT generally gives a good idea of what a document is about. But the language generation can be awkward.

- *Gisting* can be valuable: generate a rough translation to decide whether a manual translation is worthwhile.

- Automatic MT systems can also be used as a starting point for manual translation.