Language analysis and translation

- History of machine translation
- Sentence structures and analysis
- Translation levels: direct, transfer-based, interlingua
- Translation techniques: statistical, rule-based, hybrid
History of machine translation

• Initial work in 1950’s, Russian to English translation, motivated by espionage needs.

• Lack of progress in 60’s and 70’s due to problems of linguistic ambiguity and computational limitations.

• 1980’s: rule-based and statistical approaches developed.

• 90’s: widespread use of MT, including internet. Speech translation introduced.

• Recent work mainly on statistical, hybrid, example-based approaches.
Sentence structures and analysis

Chomsky’s grammars for natural language: each grammar consists of a formal set of allowed structures for sentences and phrases.

Can be expressed by grammar rules or trees. Eg:

\[
\begin{align*}
S & ::= NP + VP \\
VP & ::= Verb + NP \\
NP & ::= Det + N \\
Verb & ::= Aux + V \\
Det & ::= the, a, ... \\
N & ::= man, ball, ... \\
Aux & ::= will, can ... \\
V & ::= hit, see, ...
\end{align*}
\]

“A sentence consists of a noun phrase followed by a verb phrase. A verb phrase consists of a verb followed by a noun phrase ...”
This grammar allows construction of sentences $S$ such as “the man will hit the ball”, which has the structure (noun phrase) (verb) (noun phrase) as (the man) (will hit) (the ball).

Likewise “a ball can see a man”. Generally will be unlimited number of valid sentences for a grammar. The individual sentences can be viewed as tree structures of objects.
Sentence structure
A more comprehensive English grammar for simple sentences:

\[
\begin{align*}
S & ::= \text{NP} + \text{PRED} \\
\text{PRED} & ::= \text{VP} \mid \text{VP} + \text{NP} \\
\text{NP} & ::= \text{Prep} + \text{NP} \mid \text{Det} + \text{AP} \mid \text{AP} \\
\text{AP} & ::= \text{Adj} + \text{AP} \mid \text{Noun} \\
\text{VP} & ::= \text{Adv} + \text{VP} \mid \text{VP} + \text{Adv} \mid \text{Aux} + \text{VP} \mid \text{Verb} \\
\text{Det} & ::= \text{the, a, ...} \\
\text{Noun} & ::= \text{house, I, ...} \\
\text{Aux} & ::= \text{will, can ...} \\
\text{Verb} & ::= \text{buy, see, ...} \\
\text{Adv} & ::= \text{sometimes, often, ...}
\end{align*}
\]

Alternatives expressed with |.
Sentence analysis

- tokenising: breaking sentence into individual words and symbols

- morphological analysis: identifying the categories of individual words (as nouns, adjectives, etc) and their internal structure: gender, person (verb version), tense, etc.

- grammatical analysis: identifying structure of complete sentence.

May be ambiguities for individual words (eg: “flies”) and sentences (“the French restaurant manager”).
**Sentence analysis**

“The large white houses” tokenised into 4 words, identified from dictionary as article, adjective, adjective, noun. Noun is identified as plural form. Sentence structure identified as NP.

In other languages, adjectives vary with number and gender of noun. Eg, Russian adjective for ‘large’:

- masculine: bolshoy
- feminine: bolshaya
- neuter: bolshoe
- plural: bolshie

Different forms generated from stem “bolsh”. Likewise for ‘white’ (beliy).
Translation levels

- Direct translation: simple word-by-word mapping of source text to target language, with minimal analysis.

- Transfer-based translation: analyse input to intermediate representation, such as syntax trees, map to target language representation, then generate target text.

- Interlingua translation: analyses source to an independent semantic representation (interlingua language), from which output in different target languages can be generated.
Comparison

- Direct translation ignores role that words play in a sentence, and context. May mis-translate in cases of ambiguity. Eg, “svecha” (candle) in domestic context mis-translated as “sparkplug”.

- Transfer approach specific to one pair of languages.

- Interlingua approach permits multiple languages to be translated, but requires more semantic analysis to construct intermediate representation. Natural language understanding is unsolved AI problem.
Translation techniques

- **Rule-based**: define precise rules mapping source sentences with specific structures systematically to corresponding target sentences.

- **Statistical**: use large corpus of existing bilingual data to lookup and map sentences. (Eg.: Google translate)

- **Hybrid**: use statistical approach to improve rule-based, eg., to handle ambiguous cases.
Rule-based translation

- Also known as ‘Knowledge-based machine translation’. Use linguistic knowledge of source and target languages, encoded in analysis, grammar and translation rules, to recognise elements and structure of input sentences + map to target sentences.
- Assumes input has correct formal structure + literal meaning.
- Not so effective for idiomatic and casual language. Eg: “out of sight, out of mind” $\rightarrow$ “invisible idiot”.

Translation as transformation

Rule-based translation can be expressed as transformation of analysed sentences.

Eg.: “from the house” has structure Prep + Det + Noun.

Translation to Russian omits Det (if it is “the”, “a” or “an”), translates Prep and Noun, to RPrep, RNoun, then modifies RNoun to case required by RPrep, eg., genitive case for “from”.

Result is RPrep + genitive(RNoun).
Translation of sentence structure
Likewise, translating “the white house” to Spanish, structure is

\[ \text{Det} + \text{Adj} + \text{Noun} \]

Translation needs to map Det and Adj to correct gender versions based on gender of noun (feminine), and re-order sentence:

\[ \text{fem} (S\text{Det}) + S\text{Noun} + \text{fem} (S\text{Adj}) \]
Translation of sentence structure
Translation rules

A translation rule defines general mapping from one grammar to another.

Eg: AP to RAP:

\[
\begin{align*}
n : \text{Noun} & \mapsto rn : \text{RNoun} \\
\text{when } (n, rn) & \in \text{dictionary}
\end{align*}
\]

\[
\begin{align*}
a + ap : \text{Adj} + \text{AP} & \mapsto gen(ra) + rap : \text{RAdj} + \text{RAP} \\
\text{when } (a, ra) & \in \text{dictionary}
\end{align*}
\]

\[
\begin{align*}
ap & \mapsto rap \\
\text{gen} & = \text{gender}(rap)
\end{align*}
\]

So “large white houses” translates to “bolshie belie doma”.
Problems with rule-based translation

Many rules are needed, at least one per syntactic category (AP, VP, etc).

May be ambiguity in word analysis and translation, hence several possible translation results. Eg: “spring” has several meanings and translations. “You” in English has separate familiar/polite versions in many languages.

Input analysis needs to unambiguously classify words into categories Noun, Adj, etc, and recognise sentence structure.
Statistical translation

Given a large corpus of correlated text in two languages, translate a new word or phrase by finding the matching text of maximum probability.

Eg., if corpus had occurrences

<table>
<thead>
<tr>
<th>Russian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>svecha</td>
<td>sparkplug</td>
</tr>
<tr>
<td>svecha</td>
<td>sparkplug</td>
</tr>
<tr>
<td>svecha</td>
<td>candle</td>
</tr>
</tbody>
</table>

maximum probability (0.67) is for translation of “svecha” as “sparkplug”, which is then chosen.
Statistical translation

- Word-based statistical translation is insufficient, has been generalised to phrases.

- Avoids cost of developing grammar rules, but requires access to corpus.

- The corpus influences the translation (e.g., “hear” mapped to “bravo” based on Canadian Hansard).

- No explicit ‘understanding’ of source, so result may be nonsensical.
Hybrid translation

- Combines rule-based and statistical approaches
- Statistical analysis used to correct rule-based mapping in cases of ambiguity
Comparison of approaches

- The central problems in translation (human or machine) are the multiple interpretations (ambiguity) of source text, and multiple possible translations of a text.

- Can be addressed by ‘deep’ understanding of text, or ‘shallow’ statistical correlations of texts.

- Statistical approaches have been more successful than knowledge-based within the domains of specific corpora.

- Probably a hybrid approach is essential for further progress.