Machine Translation Survey: Introduction, Problems of Machine Translation, Is Machine Translation Possible, Brief History, Possible Approaches, Current Status.

Anusaraka or Language Accessor: Background, Cutting the Gordian Knot, The Problem, Structure of Anusaraka System, User Interface, Linguistic Area, Giving up Agreement in Anusaraka Output, Language Bridges.

Multilingual Information Retrieval - Introduction, Document Pre-processing, Monolingual Information Retrieval, CLIR, MLIR, Evaluation in Information Retrieval, Tools, Software and Resources.

Multilingual Automatic Summarization - Introduction, Approaches to Summarization, Evaluation, How to Build a Summarizer, Competitions and Datasets.

---

[1]Machine Translation Survey:

Introduction

A computer system and out comes its translation in Hindi, Tamil, English and other languages. It is inexpensive, immediate and simultaneous.

The language barriers melt away. The richness of other literatures opens up to everyone.

The world is intellectually and culturally united into one. This is the dream of people working in a fascinating area of research called Machine Translation (MT).

Goal is far from realization, the first signs of limited success are apparent.

The first faltering steps in MT were taken in the 1950s and 1960s.

But at that time, formal **liguistics and artificial intelligence** had barely been born, and computer science was in its infancy. As a result, the efforts "failed" to achieve success.

[2]Problems of Machine Translation:

MT Task Difficulties
• Initially believed bilingual dictionary and sentence reordering rules required.
• Example: translating Hindi to English sentence.
raama ne kheta jotaa.

(noun) (noun) (verb)

Ram ploughed the field

(n) (v) (n)

the Hindi words are replaced by English equivalents and reordered to fol-

low the sequence non-verb-noun instead of noun-noun-verb. Unfortunately,

raama ne gaaDii jotii. Ram prepared the cart.

or

raama ne ghoDaa jotaa Ram harnessed the horse.

Thus `jotaa' can be replaced by 'plough', 'prepare' or 'harness' depending on the sense implied in the sentence. The correct sense must first be identified for each of the words before selecting the appropriate replacement.

The sentence structure must also be interpreted correctly for translation. For example, in

I saw Ramesh on the hill with the telescope.

The telescope could have been the instrument of seeing, or Ramesh could have been carrying the telescope, or it is the hill with the telescope. The three translation are different in Hindi:

> Meine duurbiina dwaaraa Ramesh ko pahaaDii par dekhaa.
>
> Meine Ramesh ko duurbiina ke saath pahaaDii par dekhaa.
>
> Meine Ramesh ko duurbiina waalii pahaaDii par dekhaa.

Hence, it would be important to identify the relationship of the telescope correctly (called the sentence structure). Such identifications might require following a paragraph and maintaining a context.

Frequently, it is important to find the referent of the pronoun. Consider the following sentences for example:

A dog saw a cow on the road.

It started barking on seeing it.

The first 'it' can refer to a dog, cow or road. If one were translating into Hindi, the gender of the verb would depend on the gender of the referent:

> vaha use dekhakara bhounkane lagaa.

If `vaha' was referring to the cow or the road, feminine would be used.

It may sound ridiculous to even consider that the road can bark, or it may seem obvious that it is the dog which barks. However, reaching such conclusions requires more than just word replacement ability.

The problems described above point to a common theme: A sentence must be "understood" (at least partially) before it can be translated.

Natural language understanding is a hard task because it requires for-mulating not only a grammar for the language but also using background knowledge including common sense knowledge. All this must be used in complex and as yet unknown ways to process a given text.

[]Is MT Possible?:

Machine Translation (MT) Applications and Challenges

• Literature, poetry, legal texts, jokes, and double meaning material are not suitable for MT due to their difficulty and potential ambiguity.
• Current task domains for MT are circulars, official communication, minutes of meetings, technical literature, and manuals.
• These texts have limited vocabulary, fixed styles, and a short lifespan, making them ideal candidates for MT.
• There are no practical automatic translation systems except one in meteorological forecasts.
• Human-aided machine translation can be a viable option, with the translator pre-editing the text, aiding the computer during translation, or post-editing the generated text in the target language.
• Most current MT systems require post-editing, with the post editor needing to know only one language.

[]Brief History:

History of Machine Translation (MT)
• Originated in the early 50s with the realization of computer translation.
• US research groups initiated translation from Russian to English, funded by defense and intelligence.
• USSR also attempted translation from English and French to Russian.
Bilingual Dictionary Lookup and MT Systems
• Early work focused on bilingual dictionary lookup.
• Developers realized need for more than bilingual dictionary lookup.
• Despite initial optimism, MT systems failed to meet promises.
• ALPAC committee in the US concluded basic research needed and MT was not feasible in the future.
MT's Death and Revival in the Late 70s

• ALPAC report led to a decline in MT efforts in the US, with funding ceased, research groups disintegrated, and the field gaining disrepute.
• Despite this, a few research groups remained active.
• The field revived in the late 70s with the TAUM-METEO system in Canada, translating Canadian weather forecasts from English to French.
• Other systems like Titus and CULT were developed around this time.
In the 80s, the Japanese successfully completed a national project (Mu) on MT between English and Japanese. The European Community has also undertaken an ambitious project called Eurotra covering all the languages of the Community. Work has also been undertaken by groups in France, Germany, Switzerland, the US and India.

[] Possible Approaches:

There are three major types of MT systems that one can imagine based on their dependence on languages.

The first type of MT systems are designed for a particular pair of languages. If translation capability is needed between another pair of languages, a new system must be constructed.

Because of its dependence on the languages, the system has the advantage that **special features** and **similarities between the concerned languages** can be made use of. It is called the direct approach.

Interlingua-Based System Overview
• Analyzes sentence in source language and represents it in an intermediate language.
• The intermediate language is typically a formal "mathematical" language.
• A generator generates a sentence in the target language using the intermediate representation.
• Advantage: The analyzer for the source language is independent of the generator for the target language.
• System with translation capability requires only 15 parsers and generators, compared to 210 systems in the first approach.
Language Transfer Approaches
• Difficulty in defining interlingua and leveraging language similarities.
• Transfer approach is used as an intermediate method.
• Parser produces source language representation, then transfers to target language.
• Generator takes over from this intermediate approach.
• For 15 languages, 15 parsers and generators are needed, but 210 transfer components are needed.
Indian Language Context
• Large number of languages (approximately 15 major ones).
• Interlingua approach preferred.
• Nature of interlingua is open.

[] Current Status:

European Language Research Activity
• Japanese National Project (Mu) completed, resulting in an industrial prototype.
• Major development effort underway.
• Eurotra project in Europe, involving hundreds of researchers.
• Uses transfer approach for tree representation of source languages.
• Other research groups active in Europe.

India's Multilingual Resources and Social Semantics

• India has active multilingual groups, with the earliest published work by Chakraborty in 1966.
• Recent research has been conducted at Tamil University Thanjavur, National Centre for Software Technology (NCST) Bombay, Centre for Development of Advanced Computing (C-DAC) Pune, and IIT Kanpur.

• Thanjavur's research group attempted Russian to Tamil translation in 1985, with the first system translating simple sentences.
• NCST's group worked on English to Hindi translation, but no working system developed.
• C-DAC's group focuses on processing Sanskrit.
• IIT Kanpur's Akshar Bharati group applies principles of Indian traditional grammar.
• The thesis proposes exploring new types of multilingual resources using Social Semantics, a form of aggregated knowledge from millions of users.
• Social Semantics offers advantages over traditional multilingual language resources, including coverage of many languages and domains and constant growth.

The questions remains how resources of "Social Semantics for Multilingual Retrieval"
• Exploits Social Semantics resources for multilingual retrieval.
• Shows collaboratively created datasets potential for IR application.

Multilingual Retrieval Scenarios and Semantics
• Defines multilingual retrieval scenarios.
• Discusses the use of semantics in the thesis.
• Defines IR, specifically cross-lingual and multilingual retrieval.
• Presents main research questions and contributions.
• Provides an overview of all chapters guiding the thesis content.