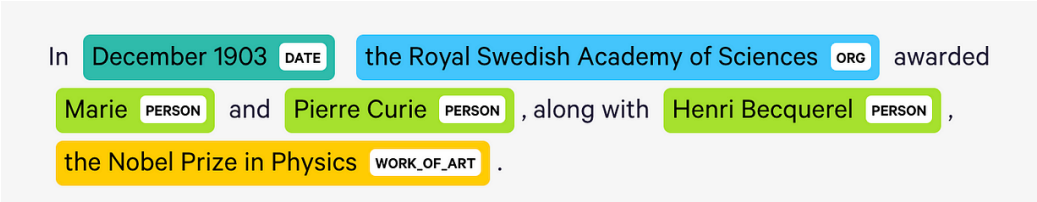
UNIT 5 : INFORMATION EXTRACTION

NAMED ENTITY RECOGNITION : **Source : Open URL**

- Named Entity Recognition (NER) is the process of identifying word or phrase spans in unstructured text (the entity) and classifying them as belonging to a particular class (the entity type).
- Both NER and Relation Extraction are foundational for many Natural Language Processing (NLP) tasks.
- For Example :

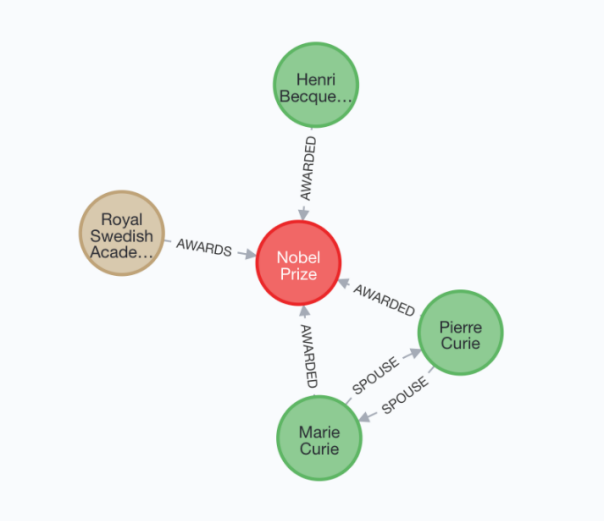


  - ❖ In the above Figure, we can see the NER process has detected that Marie Curie, Pierre Curie, and Henri Becquerel are entities of type PERSON, the Royal Swedish Academy of Sciences is an entity of type ORGANISATION, and the Nobel Prize in Physics is an object (but misclassified as a WORK_OF_ART).
- The NER process has many applications in NLP and search. For example,
  - ❖ Identifying entities in search applications means that the user can now search for "things, not strings," and thereby the user can get more relevant search results.
  - ❖ NER can be applied to large collection of documents.
  - ❖ NER can help reduce the unstructured text to a list of entities and their counts, thereby allowing clustering or summarization to help humans understand their contents.
  - ❖ The output of NER can also be input into other processes, for example, a coreference resolver that disambiguates the span "Marie" in "Marie and Pierre Curie" to entity "Marie Curie."
- Pretrained models in Named Entity Recognition (NER) are models that have been trained on large amounts of text data to recognize and extract entities such as person names, organizations, locations, and other entities from text.

- These models are usually trained using deep learning algorithms, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformers, on large datasets.
- Some popular pretrained NER models include the Stanford NER model, the spaCy NER model, and the BERT-based NER models such as BioBERT, SciBERT, and ClinicalBERT.
- Using a pretrained NER model can save time and effort in building an NER system from scratch, and can also improve the accuracy of the system, as the model has already learned to recognize and extract entities from text but performance of a pretrained NER model may vary depending on the specific domain and text data being analyzed.

RELATION EXTRACTION : **Source : Open URL**

- Relation Extraction (RE) is the process of identifying relationships implied in the text between a pair of entities.
- Both NER and Relation Extraction are foundational for many Natural Language Processing (NLP) tasks.
- RE process extracts structured information from unstructured text.
- Unstructured text means identifying the types of relationships that exist between entities, such as "works-for", "married-to", "born-in", etc.
- When applied over a large collection of text that has gone through the NER process, the RE process can extract graphs (called Knowledge Graphs) but of much greater complexity.
- Knowledge Graphs can be used for querying directly, reasoning about the entities in the graph, and to power recommendation and exploration engines.
- Example for Relation Extraction :



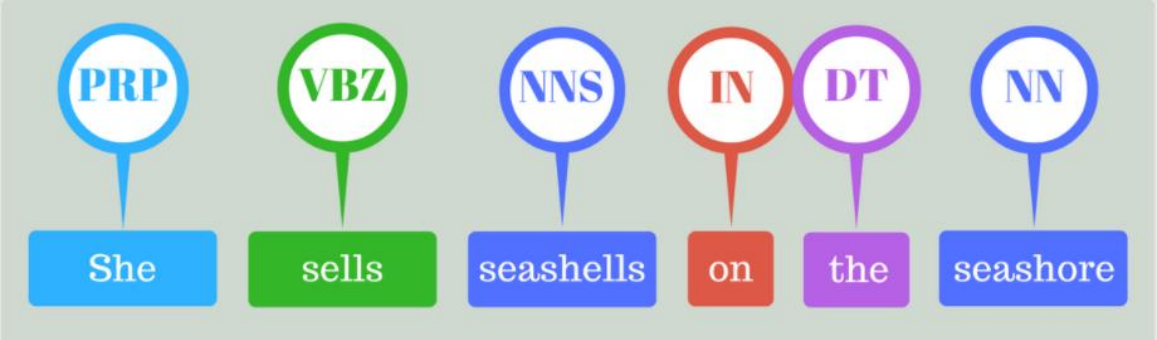
- Some of the applications are :
    - ❖ Extracting information from news articles, social media, and other sources of unstructured text.
    - ❖ Used in fields such as biomedical research, where identifying relationships between genes, proteins, and diseases is important for drug discovery and development.
- Deep learning has become an increasingly popular approach for relation extraction in natural language processing.
- Deep learning models are designed to learn multiple levels of representation from data, which can make them particularly effective for tasks that require understanding of complex relationships between entities.
- Deep learning models can be used to learn representations of words, phrases, and sentences that capture the semantic and syntactic information relevant to identifying relationships between entities.
- These representations can then be used to predict the type of relationship between entities in new, unseen text.
- CNNs can be used to extract features from different parts of a sentence, such as the subject, object, and verb, and combine them to make a prediction about the relationship between entities.
- Another common approach is to use recurrent neural networks (RNNs), which are designed to capture sequential information from data.
- RNNs can be used to model the context surrounding entities in a sentence, allowing the model to take into account the order in which words appear and their relationship to each other.

INFORMATION EXRACTION USING SEQUENTIAL LABELLING

Source : Open URL

- Sequence labelling is a Natural Language Processing task and has been one of the most discussed topics in Linguistics and Computational Linguistics history.
- Sequence labeling can also be used for tasks such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and chunking.
- Information Extraction (IE) using sequence labeling is a popular approach in natural language processing that involves tagging or labeling individual words or tokens in a text sequence with different labels or tags that represent specific information.

- In the context of IE, sequence labeling can be used to identify and extract specific types of information from text.
- It aims to classify each token (word) in a class space C. This classification approach can be independent (each word is treated independently) or dependent (each word is dependent on other words).
- For example, in Named Entity Recognition, each word in a sentence is labeled as representing a person, location, organization, or other entity. In relation extraction, each word or phrase in a sentence is labeled with a tag that indicates its role in the relationship between two entities.
- One of the most widely used sequence labeling algorithms in NLP is the Hidden Markov Model (HMM), which models the probability distribution of sequences of labels given a sequence of observations.
- Another popular approach is Conditional Random Fields (CRFs), which model the conditional probability of label sequences given input sequences.
- Deep learning models such as Recurrent Neural Networks (RNNs) and Transformers can also be used to get promising results for sequence labeling tasks in NLP, especially in Named Entity Recognition.
- These models learn to capture the contextual information in the text sequence by modeling the dependencies between each word or token and its surrounding context.
- The main advantage of Sequence labeling is, its algorithms provide a structured representation of text data that can be further processed and analyzed as per requirement.
- For example, Part-Of-Speech Tagging (POS Tagging) is a sequence labeling task. It is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech (syntactic tag), based on both its definition and its context. POS Tagging is a helper task for many tasks about NLP: Word Sense Disambiguation, Dependency Parsing, etc.



ISSUES IN MACHINE TRANSLATON

Source : Open URL

1. Translation is not straightforward :

- It means the machine translation is not replacing words for words.
- Incorrect word orders.
- Rewriting of text into another Language.
- Unable to Choose right and correct words for translation.
- For ex; Imperative mood in English , infinitive in French

2. Automation of translation is not easy :

- Translation Quality is poor .
- For ex; “fan” can be a ventilator or an enthusiast.
- Different word classes: “love” can be both verb and a noun.
- Idioms : ex: “country music “ meaning type of music.
- Personal pronouns : It means second person pronouns may vary in familiar and formal situations.

3. Issues in Morphological analysis :

- For ex ; Chinese and Japanese do not use punctuations, it means sentences are not separated by anything .

4. Issues in Syntactic Analysis :

- These are the modifiers in a problem.
- For ex; “ The boy saw a girl with a telescope “, means the girl had a telescope vs the boy used a telescope to see the girl

5. Analysis of Context :

- 20-40 words in a sentence leads to 100 million possible translations

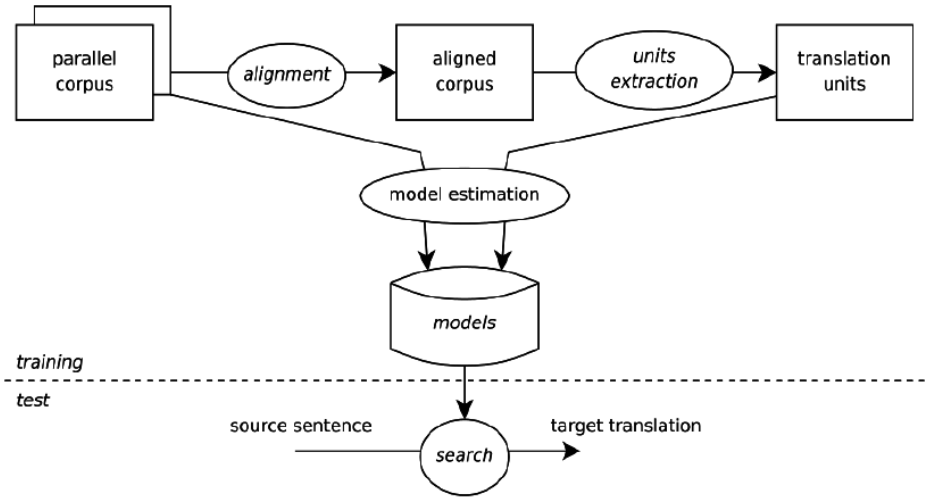
STATISTICAL MACHINE TRANSLATON

Source : Open URL

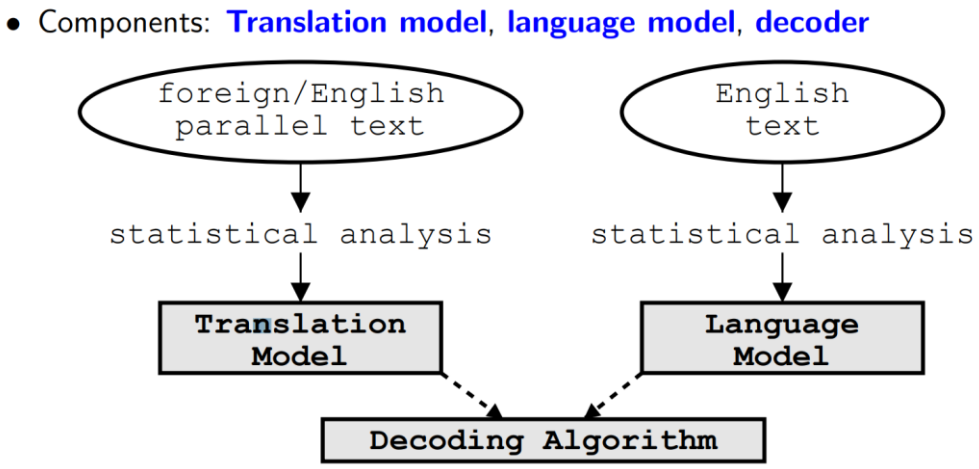
- Statistical Machine Translation (SMT) learns how to translate by analyzing existing human translations (known as bilingual text corpora).

- In contrast to the Rules-Based Machine Translation (RBMT) approach that is usually word-based, most modern SMT systems are phrase-based and assemble translations using overlap phrases.
- In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called phrases, but typically are not linguistic phrases, but phrases found using statistical methods from bilingual text corpora.
- Analysis of bilingual text corpora (source and target languages) and monolingual corpora (target language) generates statistical models that transform text from one language to another with that statistical weights are used to decide the most likely translation.
- Statistical machine translation (SMT) deals with automatically mapping sentences in one human language (for example, French) into another human language (such as English). The first language is called the source and the second language is called the target. This process can be thought of as a stochastic process.
- There are many SMT variants, depending upon how translation is modeled. Some approaches are in terms of a string-to-string mapping, some use trees-to-strings, and some use tree-to-tree models. All share in common the central idea that translation is automatic, with models estimated from parallel corpora (source-target pairs) and also from monolingual corpora.
- SMT is based on the idea that the probability of a given translation is dependent on the probability of the source sentence and the target sentence.
- SMT works by analysing a large corpus of parallel texts, which are pairs of sentences in the source and target languages that have been professionally translated. From this corpus, SMT builds a statistical model that captures the probability of different translations for a given sentence.
- The SMT system is then trained on this data to estimate the most likely translation of an input sentence.
- The system does this by breaking down the input sentence into smaller units, such as words or phrases, and then using the statistical model to estimate the probability of different translations for each unit. The system then combines these translations to produce a full translation of the input sentence.
- SMT has several advantages over other types of machine translation:
    - ❖ Rule-based machine translation.
    - ❖ SMT can be trained on a large amount of parallel text, which allows it to capture the nuances and idiomatic expressions of a language.

- ❖ SMT can be adapted to different domains or subject areas by training it on specialized parallel text in those areas.
- ❖ SMT can continuously learn and improve its translations based on new data and feedback.
- ❖ SMT is scalable, meaning it can easily handle large volumes of text and multiple languages simultaneously.
- ❖ Compared to traditional human translation services, SMT is relatively low cost and can provide translations at a faster rate.

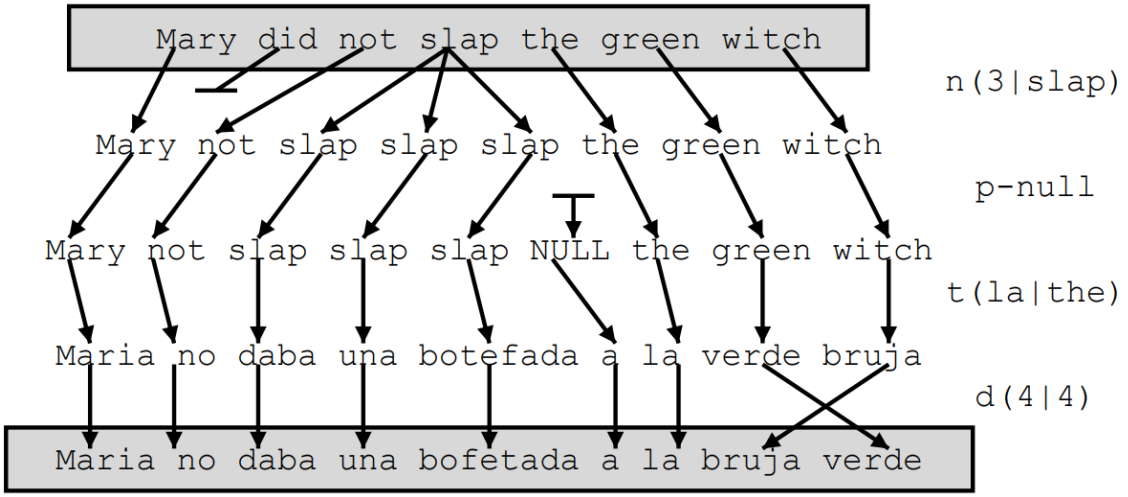


Components of Statistical Machine Translation : **Source : Open URL**



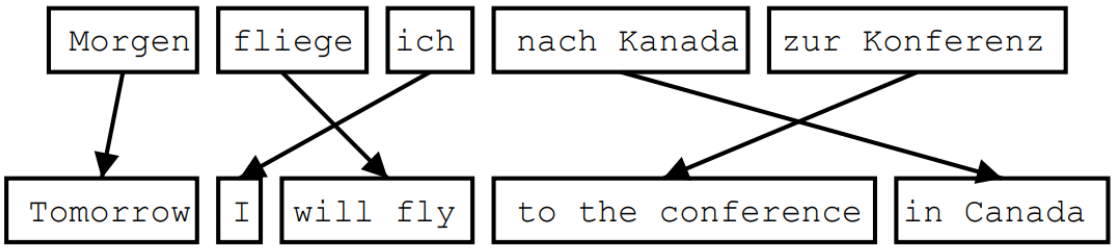
Types of SMT approaches :

1. Word-based models



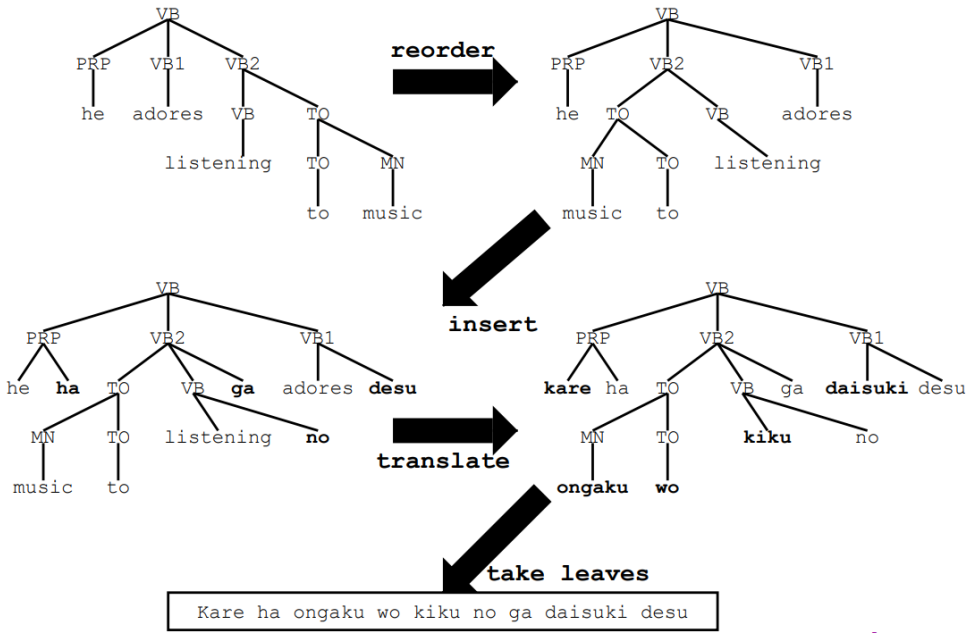
- A word-based model is a type of Statistical Machine Translation (SMT) model that focuses on translating individual words from the source language to the target language.
- In this model, each word in the source sentence is translated independently, without considering the context of the surrounding words.
- The target sentence is then generated by combining the translated words in the correct order.
- One of the main challenges with word-based models is handling words that have multiple meanings or that are ambiguous in context.
- In these cases, the model may select the wrong meaning of the word or produce a translation that is not appropriate for the context.

2. Phrase Based Models



- Foreign input is segmented in phrases– any sequence of words, not necessarily linguistically motivated.
- Here Each phrase is translated into English .
- This is the most widely used SMT model.
- It breaks down the input sentence into phrases and uses statistical models to translate these phrases.
- These models typically use phrase-to-phrase translation tables and language models to generate translations.

3. Syntax Based Models :

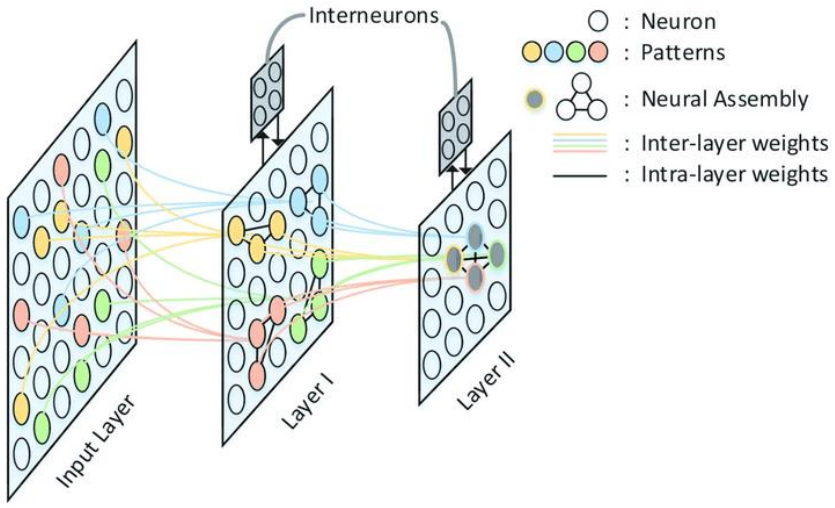


- Syntax-based models are a type of Statistical Machine Translation (SMT) model that uses the syntactic structure of the source and target languages to improve translation accuracy.
- In a syntax-based model, the input sentence is first parsed into a tree structure that represents the syntactic relationships between the words.
- This tree structure is then used to guide the translation process, with the target sentence being generated based on the corresponding syntactic structure in the target language.

4. Neural machine translation (NMT) models:

- This is a more recent type of SMT model that uses deep neural networks to learn to translate text.

- NMT models have shown to be effective in capturing long-range dependencies between words and producing fluent translations, but require large amounts of training data and computational resources.



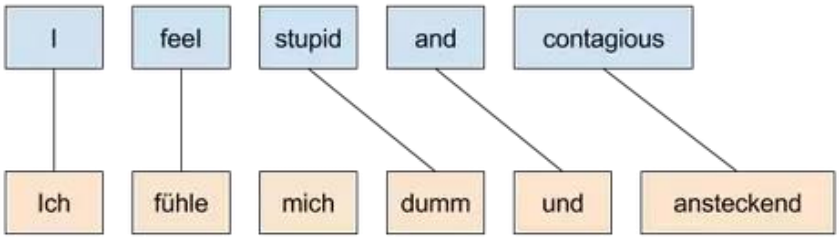
5. Hybrid models:

- These models combine the strengths of multiple SMT models to improve translation accuracy.
- For example, a hybrid model may use a phrase-based model for the initial translation, followed by a syntax-based model to improve the grammatical correctness of the output.

WORD ALIGNMENT **Source : Open URL**

- Word alignment is a technique used in tasks such as machine translation and bilingual text analysis.
- It is a natural language processing task of identifying translation relationships among the words (or more rarely multiword units) in a bitext, resulting in a bipartite graph between the two sides of the bitext, with an arc between two words if and only if they are ...
- Word alignment is an important component of a complete statistical machine translation (SMT) pipeline.
- The objective of the word alignment task is to discover the word-to-word correspondences in a sentence pair. Models for word alignment depend on the way they decompose this problem.

- Word alignment is typically done after sentence alignment has already identified pairs of sentences that are translations of one another.
- Bitext word alignment or simply word alignment is an important supporting task for most methods of statistical machine translation.
- The process of word alignment involves assigning a correspondence between each word in the source language sentence and its corresponding word in the target language sentence.
- This correspondence is typically represented using an alignment matrix, which shows the correspondence between each word in the source sentence and each word in the target sentence.



- For example ; In the German translation of the line from “Smells Like Teen Spirit” by Nirvana, the word “mich” doesn’t have a correspondent word in the original version.
- Statistical machine translation models tunes their parameters based on the word alignment.

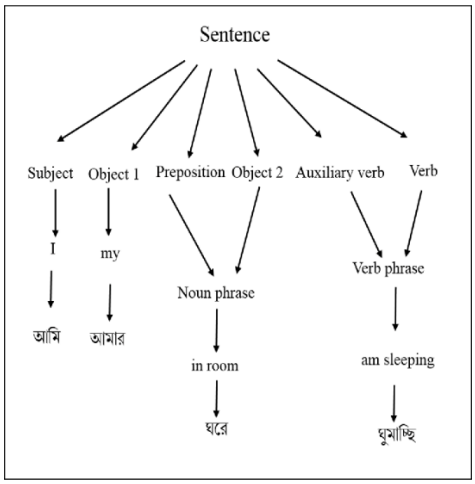
- Word alignment can be performed using different methods, such as:

1. **Manual alignment:** This involves manually identifying the corresponding words in a pair of sentences or documents. This is a time-consuming and expensive process but can produce accurate alignments.
2. **Heuristic alignment:** This involves using heuristics or rules to automatically align words in a sentence. For example, the "grow-diagonal" algorithm starts by aligning the diagonal cells of the alignment matrix and then grows the alignment in adjacent cells based on certain criteria.
3. **Statistical alignment:** This involves using statistical models to automatically align words in a sentence. For example, IBM models use statistical methods to estimate the probability of a given word alignment.

PHRASE BASED TRANSLATION MODELS

Source: Open URL

- Phrase-based translation models give much improved translations over the IBM models, and give state-of-the-art translations for many pairs of languages.
- Crucially, phrase-based translation models allow lexical entries with more than one word on either the source-language or target-language side: for example, we might have a lexical entry (le chien,the dog) specifying that the string “le chien” in French can be translated as the dog in English.
- The option of having multi-word expressions on either the source or target-language side is a significant departure from IBM models 1 and 2, which are essentially word-to-word translation models (i.e., they assume that each French word is generated from a single English word).
- Multi-word expressions are extremely useful in translation; this is the main reason for the improvements that phrase-based translation models give.
- The translation models use statistical algorithms to find the best possible translation for each phrase based on the context and the probability of occurrence of the phrase in the target language.
- Once all the phrases are translated, they are reassembled to form a complete sentence in the target language.
- This process allows for better translation accuracy since the translation model can consider the context and meaning of each phrase and provide a more fluent and accurate translation. Example;



SYNCHRONOUS GRAMMARS

Source : Open URL

- Synchronous context-free grammars are a generalization of context-free grammars (CFGs) that generate pairs of related strings instead of single strings.
- Thus they are useful in many situations where one might want to specify a recursive relationship between two languages.

- Originally, they were developed in the late 1960s for programming-language compilation.
- In natural language processing, they have been used for machine translation and (less commonly, perhaps) semantic interpretation.
- As a preview, consider the following English sentence and its (admittedly somewhat unnatural) equivalent in Japanese (with English glosses):

(1) The boy stated that the student said that the teacher danced

(2) shoonen-ga gakusei-ga sensei-ga odotta to itta to hanasita
the boy the student the teacher danced that said that stated

- One might imagine writing a finite-state transducer to perform a word-for-word translation between English and Japanese sentences, but not to perform the kind of reordering seen here. But a synchronous CFG can do this.
- In synchronous grammar, the rules are used to map a sentence from one language to a sentence in another language, using a set of alignment links that connect words and phrases in the two sentences.
- This is useful for tasks like machine translation, where the goal is to automatically translate text from one language to another.
- Synchronous grammars are typically represented as a set of production rules that define the mapping between the two languages.
- These rules can be written using formal notation like context-free grammars, and they can be applied in a bottom-up or top-down fashion to generate a translation.
- One of the key benefits of using synchronous grammars in NLP is that they can capture the complex structural relationships between languages, which can be difficult to model using traditional machine translation techniques.
- Additionally, synchronous grammars are often more compact and expressive than other types of grammars, which makes them easier to work with and analyze.

Advantages of Synchronous Grammars :

- ❖ Synchronous grammars can be able to capture complex structural relationships
- ❖ Synchronous grammars are often more compact and expressive than other types of grammars, which makes them easier to work with and analyze.
- ❖ Synchronous grammars can also be used to translate between multiple languages at once, which can be useful in multilingual applications.

- Synchronous grammars can improve the performance of machine translation systems by reducing the number of possible translations and improving the accuracy of the translations that are generated.
- Increased Flexibility

Limitations /Disadvantages of Synchronous Grammar :

- Complexity: Developing and maintaining a synchronous grammar can require significant expertise and resources.
- Limited scope: Synchronous grammars are designed to operate on pairs of sentences in different languages. This means that they may not be well-suited for tasks that require more general NLP capabilities, such as text classification or sentiment analysis.
- Limited availability: Synchronous grammars may not be widely available for all language pairs, which can limit their usefulness in multilingual applications.
- Performance: Although synchronous grammars can improve the performance of machine translation systems, they may not always produce the most accurate translations.