Applied Human Language Technology

Machine Translation

Dr. Irene Murtagh

This Week:

- Introduction and brief history
- Multilingual Computing
- Transfer and Interlingua Machine Translation (MT)
  - Transfer: Syntactic, Semantic, Lexicalist
  - Interlingua: KB and Linguistics-Based
  - Example Based MT and Translation Memory

Introduction

- Machine Translation: use of computers to translate or help translate human languages.

- Relatively long history (late 1940s) in computing terms.

- Highly determined by economic and political circumstances.

Brief History

- 1954 Georgetown IBM experiment Russian to English. Claimed MT would be resolved in 3-5 years.
- ALPAC report (1966) – concluded MT was too expensive and ineffective
- 1980s – Transfer based approaches
- 1988 – Word based models (IBM models) (initial statistical approach SMT)
- 1990 – knowledge based systems that use an interlingua representation as an intermediate step between input and output.
- 1997 – Internet takes off Babelfish was introduced, and Google Translate came to be almost 10 years later in '06.
- 2003 – Phrase based models (Philip Koehn) (next statistical approach)
- 2006 – Google Translate (and Moses the following year)

Brief History

Neural Machine Translation

- 2013 First papers are introduced on **Neural Machine Translation** (NMT).
- 2016 NMT was launched in production in companies
- 2020 Current approach focusing on neural machine translation + machine learning.

Sample of Techniques Used

- Database access (e.g. word look up).

- Parsing (e.g. source language analysis).

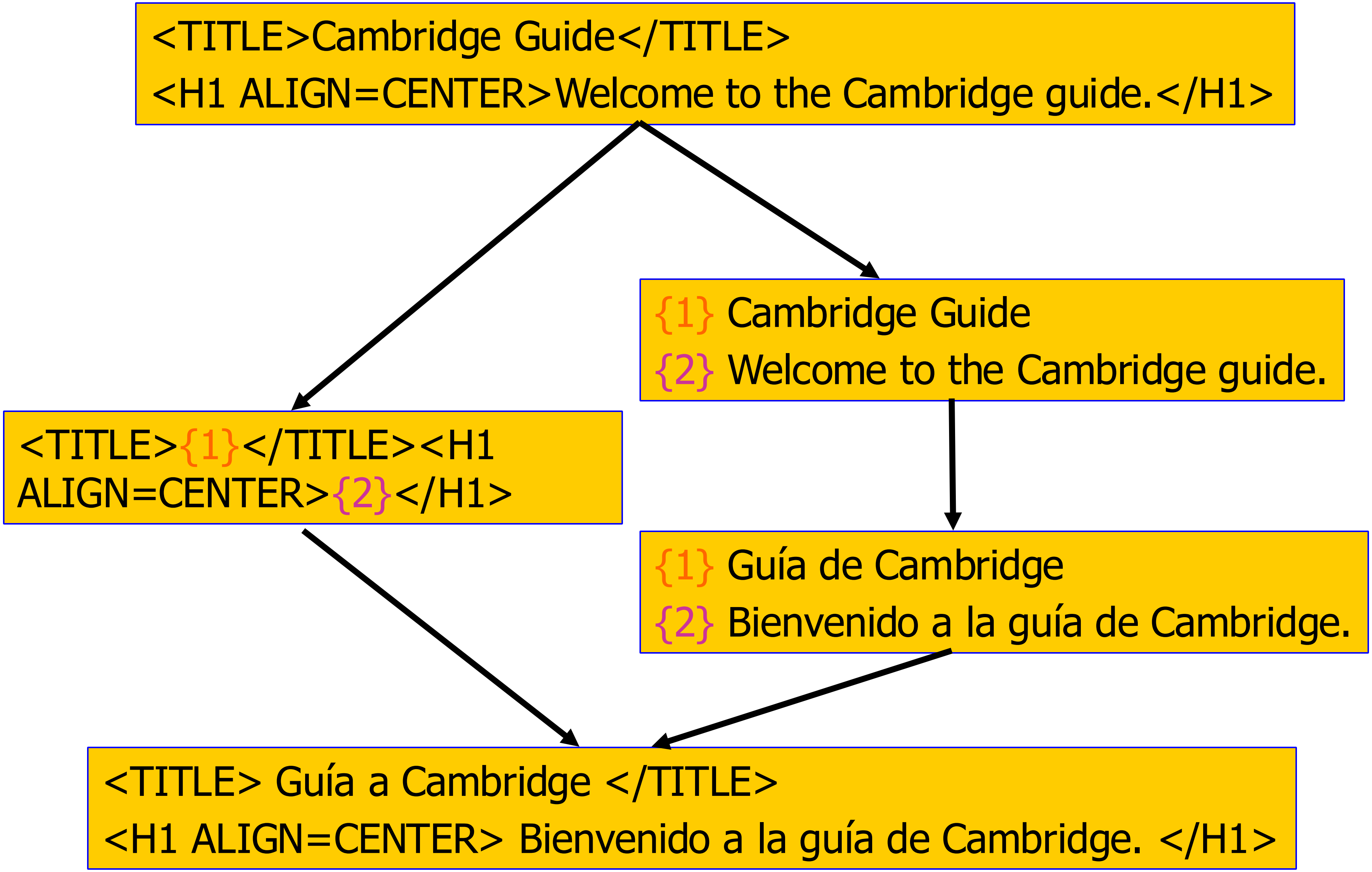
- String matching (e.g. idiom translation).

- Statistics and probability (e.g. sentence alignment).

- Knowledge representation (e.g. disambiguation).

Text Formatting

- Plain text is hardly ever used.
- HTML, Latex, RTF, MS Word, etc.
- Text must be extracted prior to translation.
- After translation, formatting must be restored.
- Useful for batch translation.



<TITLE>Cambridge Guide</TITLE>
<H1 ALIGN=CENTER>Welcome to the Cambridge guide.</H1>

{1} Cambridge Guide
{2} Welcome to the Cambridge guide.

<TITLE>{1}</TITLE><H1
ALIGN=CENTER>{2}</H1>

{1} Guía de Cambridge
{2} Bienvenido a la guía de Cambridge.

<TITLE> Guía a Cambridge </TITLE>
<H1 ALIGN=CENTER> Bienvenido a la guía de Cambridge. </H1>

Considerations: webpage translations

- **Syntax** of the formatting language; preferably from its specifications.

- Need to identify **text requiring translation**.

- **Commands** containing text to be translated?

  (e.g. <IMG ALT="Map">)

- Text may move with its **tags**:
  - She has a **<B>**red**</B>** car.
  - Tiene un coche **<B>** rojo **</B>.**

Multilingual Computing

- Need to go beyond ASCII.
- Many more fonts needed.
- Decide how text is to be entered.
- Display data (e.g. dates, currencies) in locale sensitive format.
- Use standards for representing characters.

Character Sets

- Character sets: characters used in a language, regardless of computer representation.
- Not really a question in English.
- In Chinese some subset of the tens of possible characters needs to be promoted.

Character Codes

- An assignment of numbers to each character in the character set.
- ASCII assigns 0-127,
  i.e. just under one byte needed (e.g. A=65 and z=122).
- Fixed width: same no. of bytes per character, typically 1.
- Variable width: different characters may be encoded by different no. of bytes.

Beyond ASCII

- Extensions to ASCII for European and other alphabetic scripts (e.g. Greek, Arabic).

- ISO 8859 series: these employ the unused 128 codes left by ASCII. E.g. 8859-1 (a.k.a. Latin-1) encodes ñ=241.

Representation space for a byte:

ASCII (0-127)	Extensions (128-255)

UNICODE

- Consortium standard, intended for most of the world’s scripts.

- Two bytes (65,535 codes), fixed-width.
  e.g. for Latin-1 and ASCII first byte is 0.

- Divided into areas for encoding various scripts and other symbols.

- But note: **UNICODE** is *not* a font.

UNICODE Areas

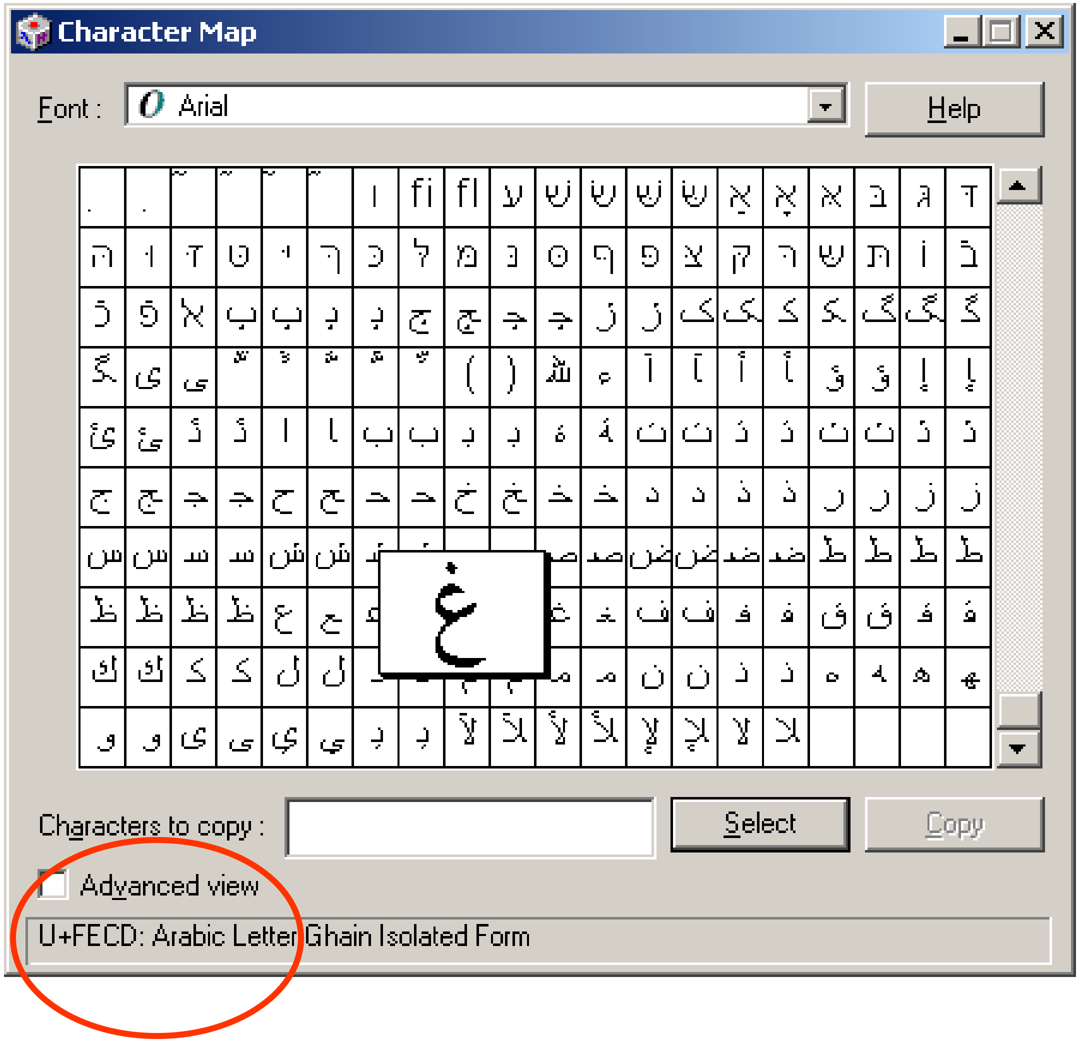
0-8,191	General scripts (e.g. English, Arabic)
8,192-10,175	Symbols (e.g. arrows)
12,288-13,311	CJK (Chinese, Japanese, Korean)phonetics/symbols
19,968-40,959	CJK ideographs (e.g. Kanji)
44,032-55,203	Hangul Syllables (for Korean)
55,296-57,343	Surrogates (for range extensions)
57,344-63,743	Private use (e.g. for companies)
63,744-65,535	Compatibility area (e.g. earlier standards)

Input Methods

- Keyboards are most common input mechanism, with about enough keys for all characters.

- Most alphabetic scripts (e.g. Latin, Arabic, Hebrew) can be accommodated easily.

- Still, many characters need more than one action (e.g. capital letters).



Character Map

Font : Arial

Characters to copy :

□ Advanced view

U+FECD: Arabic Letter Ghain Isolated Form

Multiple Action Input Methods

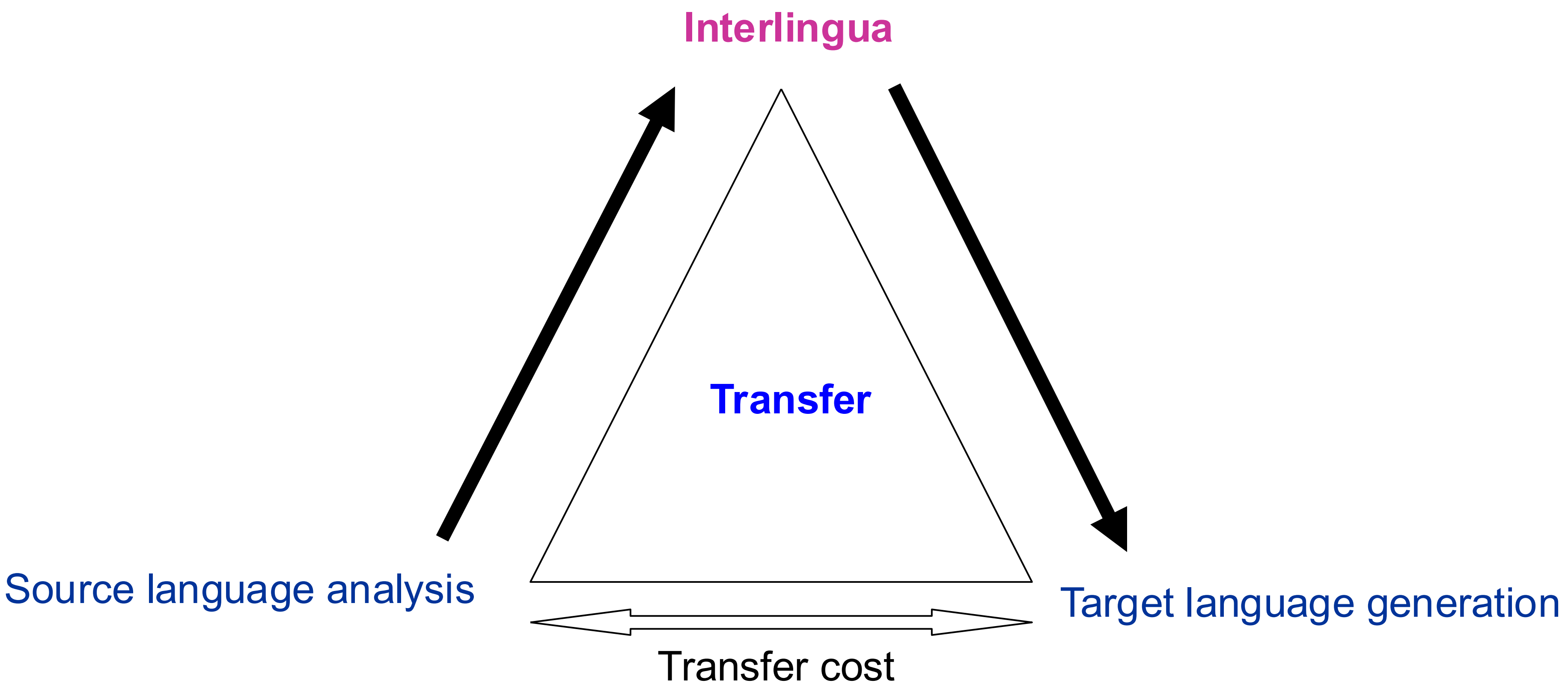
- Pressing SHIFT or ALT or CTRL (or all at once!) to get a character.

- CJK have several input methods:
  - **Pronunciation-based.**
  - **Shape-based**: enter character’s radicals.
  - **Code-based**: enter the character’s code.
  - **Association-based**: let user define key sequence.

Paradigms

- A theory or set of beliefs within which scientists work.
- 1980's and 1990's
  - Transfer vs Interlingua
    - Syntactic, Semantic and Lexicalist **Transfer**
    - Knowledge- vs Linguistics-Based **Interlingua**
  - Example-Based MT vs Translation Memory

Interlingua vs Transfer



Interlingua

Transfer

Source language analysis

Target language generation

Transfer cost

Vauquois Triangle

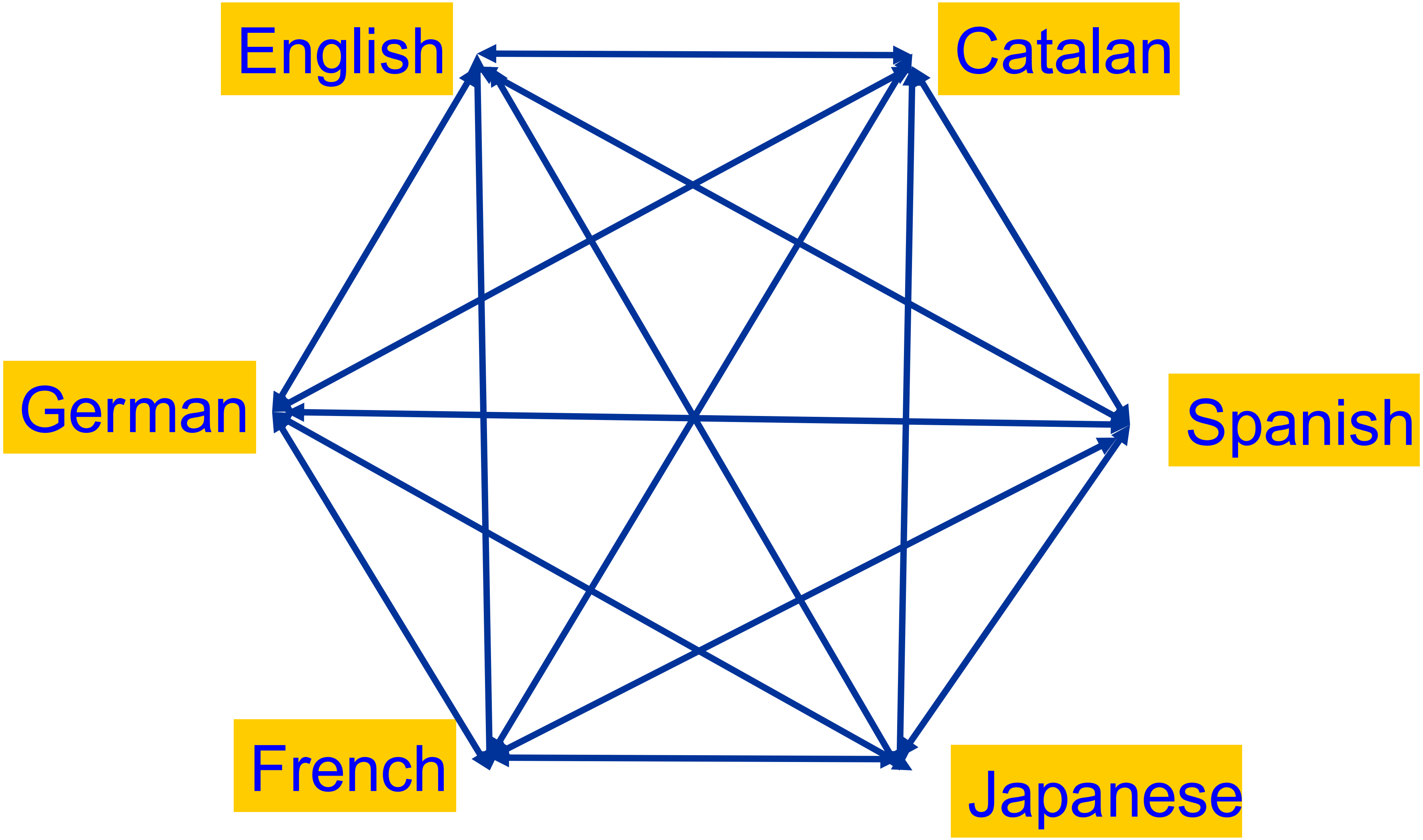
Transfer:

- Contrasts are fundamental to translation.
- Statements in one theory (source language) are mapped into statements in another theory (target language).

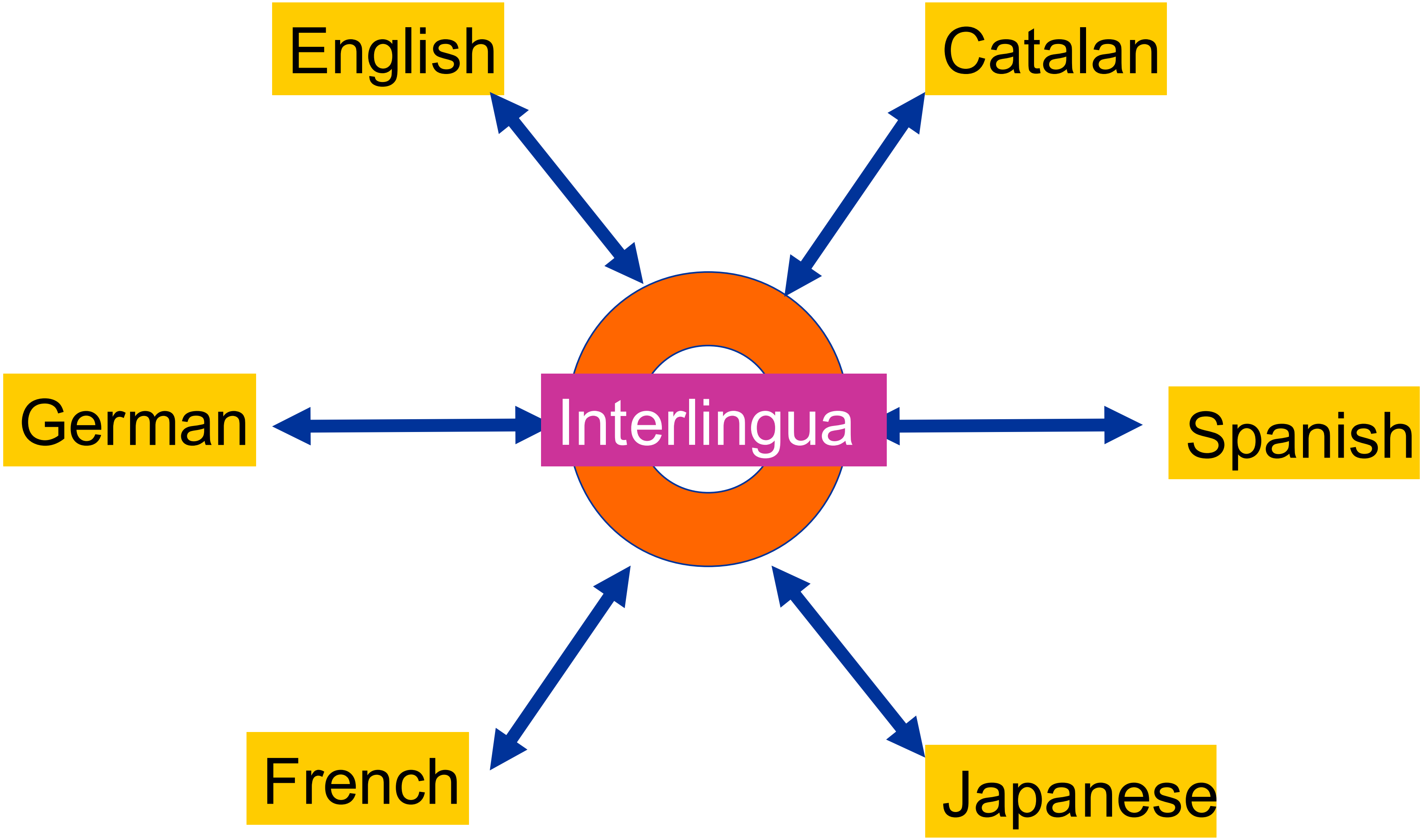
Interlingua:

- Meanings are language independent and can be encoded.
- They are extracted from **SL** (source language) sentences and rendered as **TL** (target language) sentences.

Multilinguality - Transfer



Multilinguality - Interlingua



Transfer

- Easier to implement
- Good for mono- or bi-directional systems
- Humans work on 2 languages at a time

- Modifications affect several transfer modules
- Inefficient for multilinguality

Interlingua

- Eliminates redundancy
- Highly modular
- Simplifies addition of languages

- Different linguists may disagree on representation of meaning
- Difficult to ensure that TL generator can produce sentence from SL representation

Paradigms

Paradigm : A theory or set of beliefs within which scientists work.

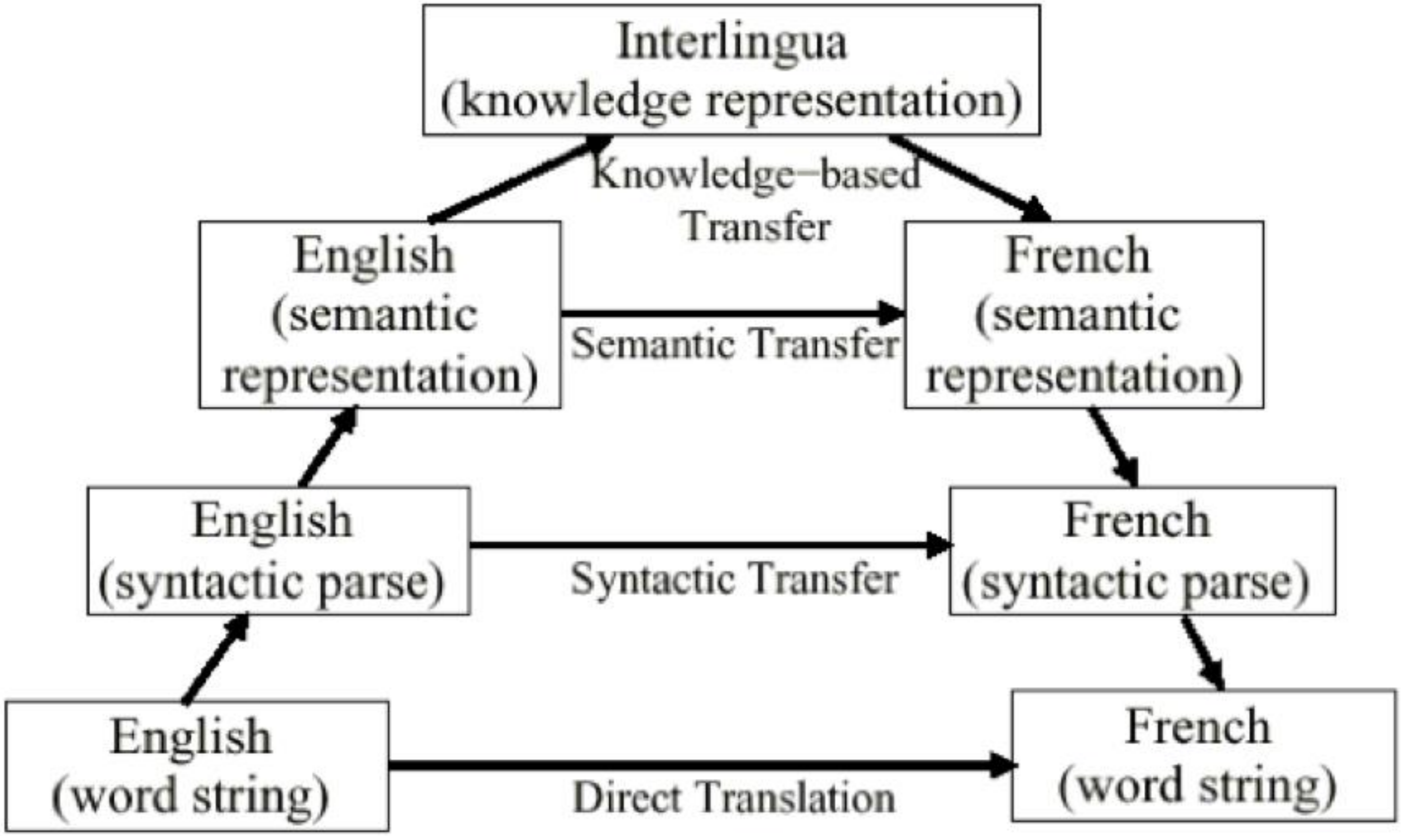
There are a number of different approaches:
Rule Based Machine Translation (RBMT)

1. Transfer vs Interlingua
2. Syntactic, Semantic and Lexicalist Transfer
3. Knowledge- vs Linguistics-Based Interlingua

Corpus Based Machine Translation (CBMT)

1. Example-Based MT vs Translation Memory
2. (Data driven) Statistical MT Approach vs. Machine Learning Neural MT Approach

Approaches to MT



Interlingua (knowledge representation)

Knowledge−based Transfer

English (semantic representation)

Semantic Transfer

French (semantic representation)

English (syntactic parse)

Syntactic Transfer

French (syntactic parse)

English (word string)

Direct Translation

French (word string)

Transfer MT

- Analysis of source language data using a morphological analyser, parser and a grammar

- Depending on approach, grammar must **build syntactic / semantic representation**

- **Transfer** to target language model.

- **Generation of target language data** output using grammar and morphological synthesiser.

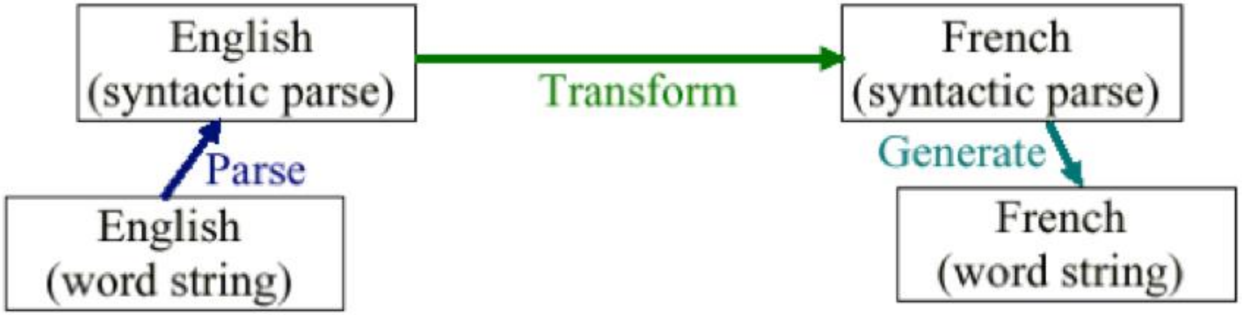
Direct Translation: Example

- Input
  - watashihatsukuenouenopenwojonniageta
- Morphological Analysis
  - watashi ha tsuke no ue no pen wo jon ni ageru PAST
- Lexical transfer of content words
  - I ha desk no ue no pen wo John ni give PAST.
- Preposition re-arrangement
  - I ha pen on desk wo John to give PAST
- SVO rearrangements & determiners
  - I give PAST the pen on the desk to John
- Morphological Generation
  - I gave the pen on the desk to John

Direct Translation

- Series of processing stages
  - Each focused on a single problem (e.g., morphological analysis)
- Stages manipulate strings of tokens
  - No parsing or syntactic structures.
- Each stage performs a uni-directional transformation on the input.

Syntactic Transfer

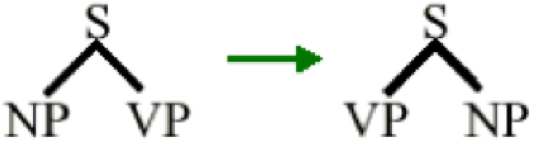


Three steps:

- **Parse the source text.**
- **Transform the source language syntax tree into the target language.**
- **Use the target language syntax tree to generate a sentence.**

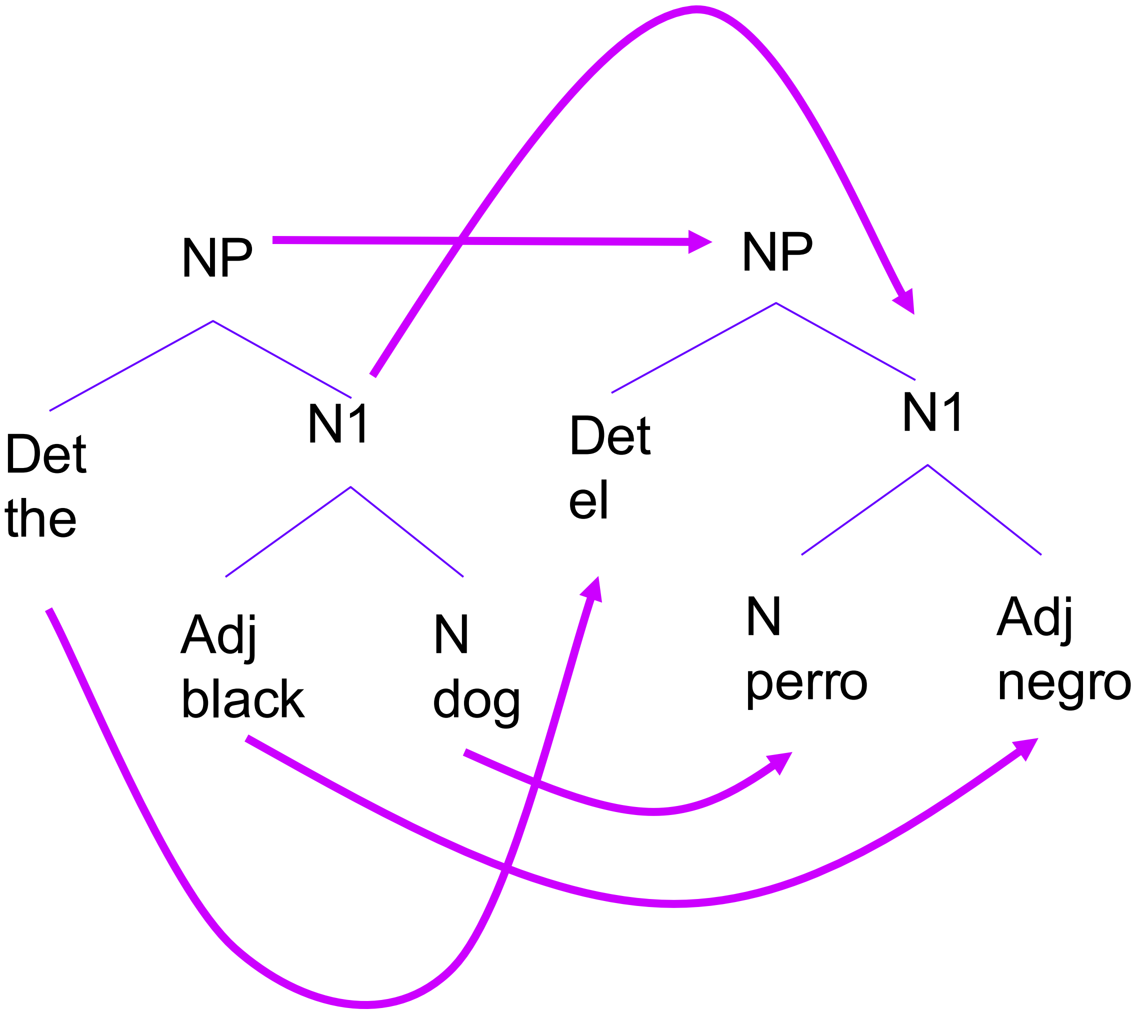
Syntactic Transfer

- Define transformational rules on syntax trees

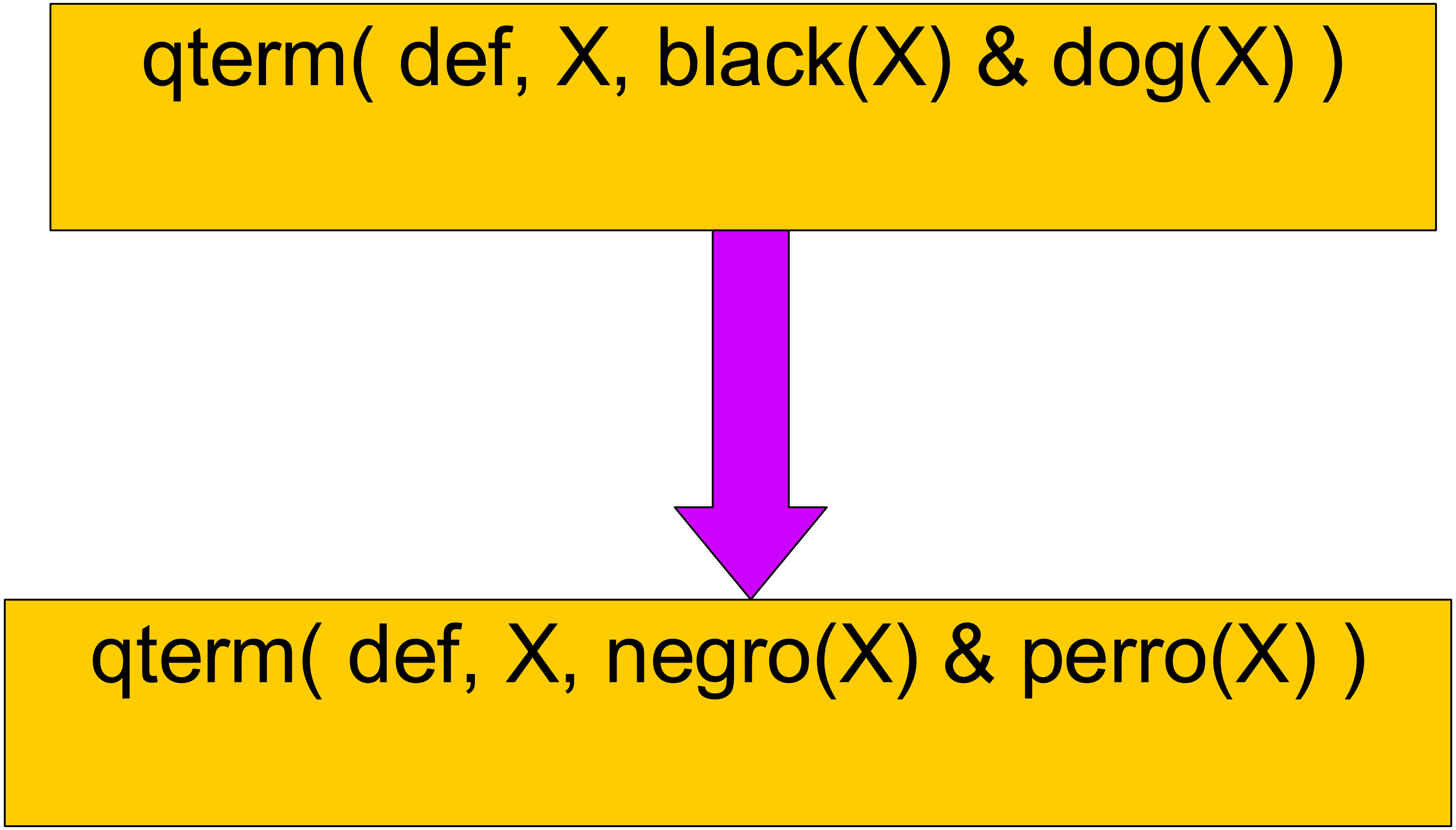


  - Context-free rules
  - Context-sensitive rules
- Apply rules to the source language syntax tree.
  - Top-down or bottom-up

Syntactic Transfer



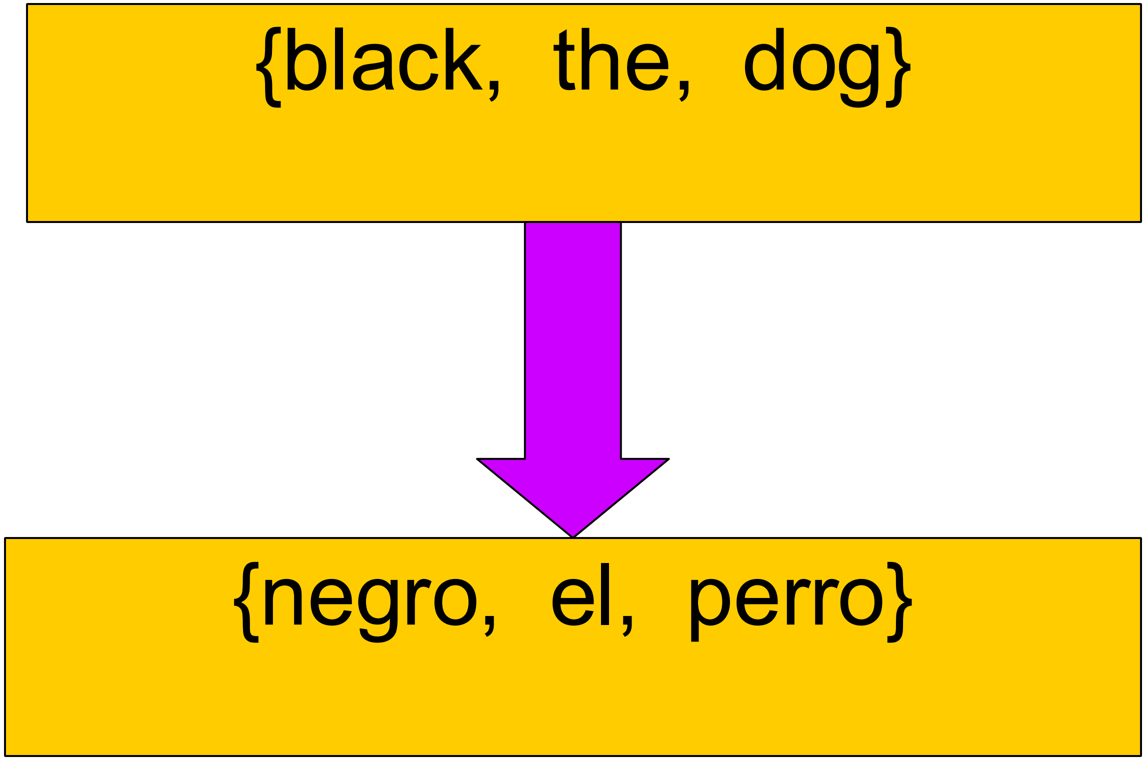
Semantic Transfer



qterm(def, X, black(X) & dog(X))

qterm(def, X, negro(X) & perro(X))

Lexicalist Transfer



{black, the, dog}

{negro, el, perro}

Syntactic:

- Rearrangement of phrases and translation of lexical items

Semantic:

- Accurate, formal and well-motivated **representations of meaning** offer the greatest chance of meaning preservation during translation.

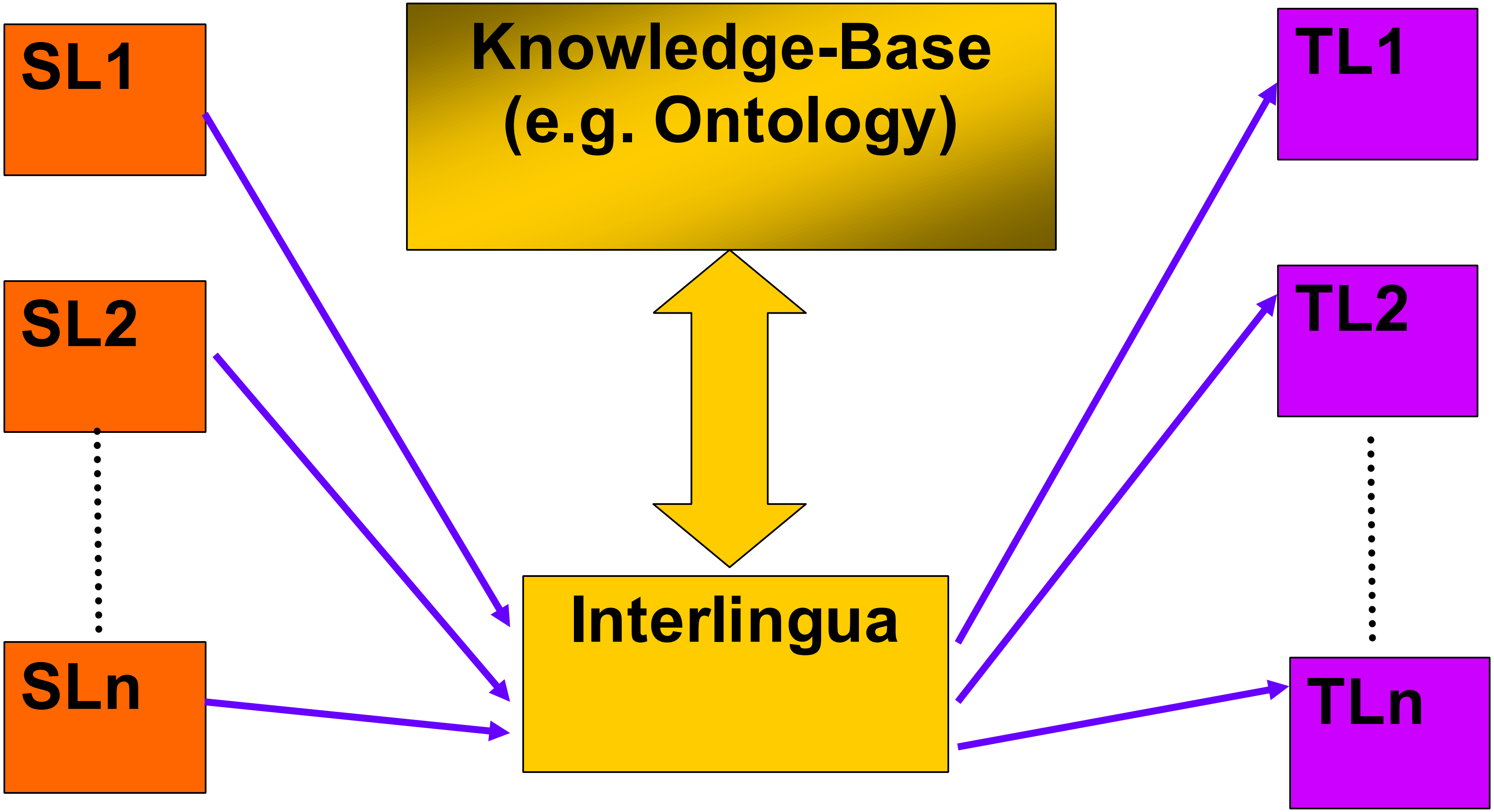
Lexicalist:

- Translation equivalences between sets of lexical items are easier to verify than more abstract representations.

Interlingua: Difficulties

- Universal lexicon
  - How do we construct a universal lexicon?
  - Must include all distinctions made by *any* language.
  - How to differentiate similar terms?
    - e.g., "shake" vs "vibrate"
- Universal knowledge format
  - How do we encode "knowledge"
  - What to include? (e.g., pragmatic information?)
- Unnecessary disambiguation
- Preserving ambiguity

Linguistics- vs Knowledge-based Interlingua



Linguistics-Based Interlingua:

- The **language faculty** constrains meaning representations.
- **Linguistic semantic theories** identify these constraints and hence provide a sufficient basis for an interlingua representation.

Knowledge-Based Interlingua:

- **Linguistic meaning** is dependent on non-linguistic knowledge.
- Use **real world knowledge** to augment meaning representations.

The move from RBMT to CBMT:

Corpus Based Machine Translation (CBMT) uses a bilingual parallel corpus to obtain knowledge for new incoming translations.

This approach uses a large amount of raw data in the form of parallel corpora.

This raw data contains text and their translations. These corpora are used for acquiring translation knowledge.

Corpus based approach can be further classified into following two sub approaches: Statistical Machine Translation and Example-based Machine Translation Approach.

Statistical Machine Translation (SMT)

Statistical machine translation (SMT) is generated on the basis of statistical models.

The parameters for the moduls are derived from the analysis of bilingual text corpora. The initial model of SMT, based on Bayes Theorem, takes the view that every sentence in one language is a possible translation of any sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system.

A document is translated according to the probability distribution function indicated by $p(e|f)$, which is the Probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL E (for example, Ibo). The problem of modeling the probability distribution $p(e|f)$ has been approached in a number of ways.

Statistical Machine Translation (SMT)

Statistical machine translation (SMT) is generated on the basis of statistical models.

The parameters for the models/engines are derived from the analysis of bilingual text corpora. The initial model of SMT, based on Bayes Theorem, takes the view that every sentence in one language is a possible translation of any sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system.

A document is translated according to the probability distribution function indicated by:

$p(e|f)$

This is the probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL E (for example, French). The problem of modelling the probability distribution $p(e|f)$ can been approached in a number of ways.

Statistical Machine Translation (SMT)

Challenges of Statistical Machine Translation Approach

- Corpus creation can be costly for users with limited resources.
- The results are unexpected. Superficial fluency can be deceiving.
- Statistical machine translation does not work well between languages that have significantly different word orders (e.g. Japanese and European languages).
- The benefits are overemphasized for European languages.

The move to Example-Based MT :

A sentence is translated by analogy, using previous translations as examples. Use parts of examples as needed.

Throughout the 1970s and 1980s, Machine Translation (MT) research focused on the development of so-called “second generation” systems, which aimed to translate text by a process of rule-driven linguistic processing, usually in three stages: syntactic, semantic analysis of the source text, bilingual transfer at a more or less abstract level of representation, and target-text generation from syntactic representation

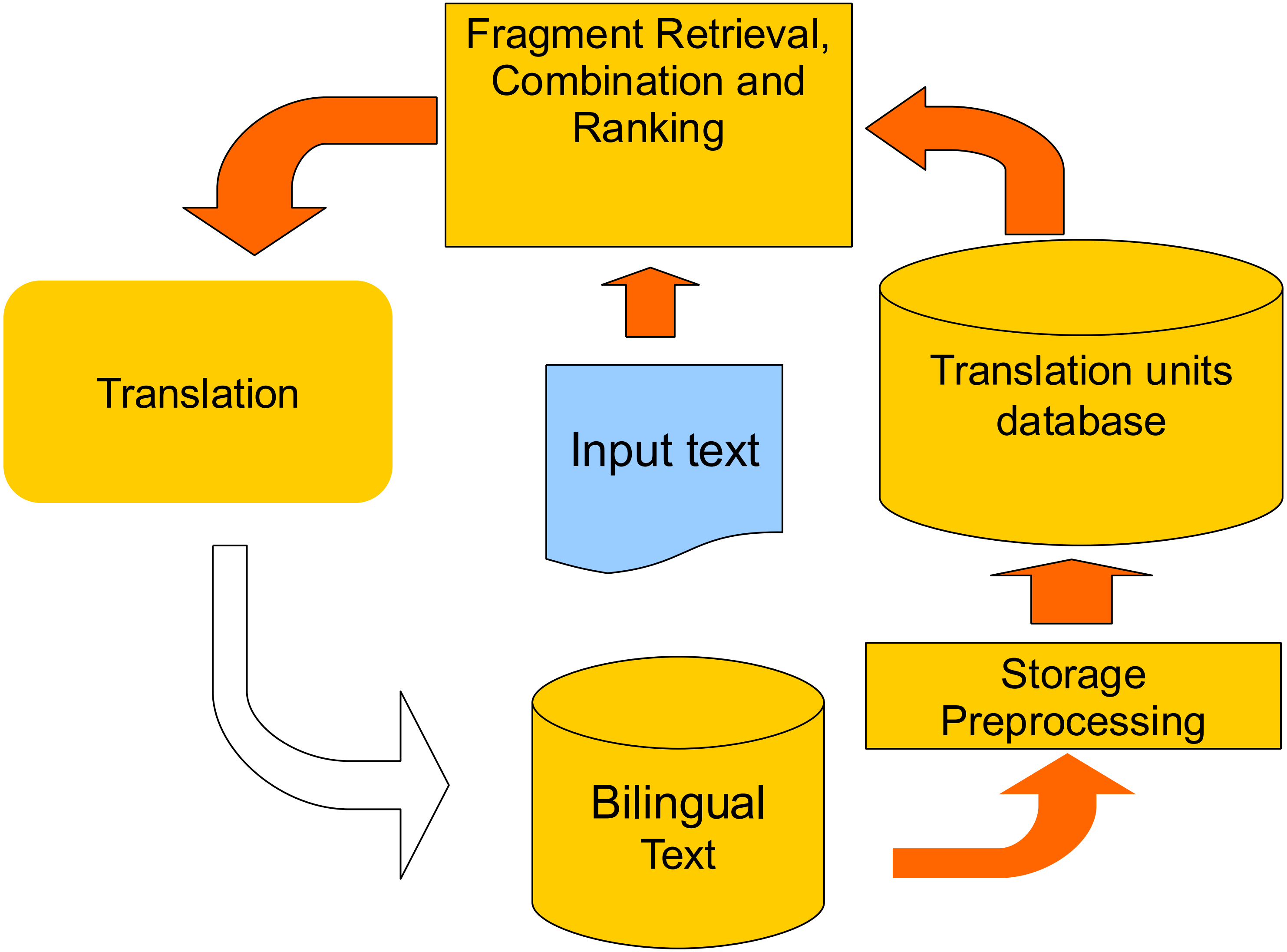
In the early 1990s, with these ideas fairly well established, and perhaps even growing stale, research in MT was hit by an apparently new paradigm in which in particular the reliance on linguistic rule systems was to be (at least partially) replaced with the use of a corpus of already-translated examples which would serve as models to the MT system on which to base its new translation. This approach came to be known as Example-Based MT (EBMT)

Translation Memory:

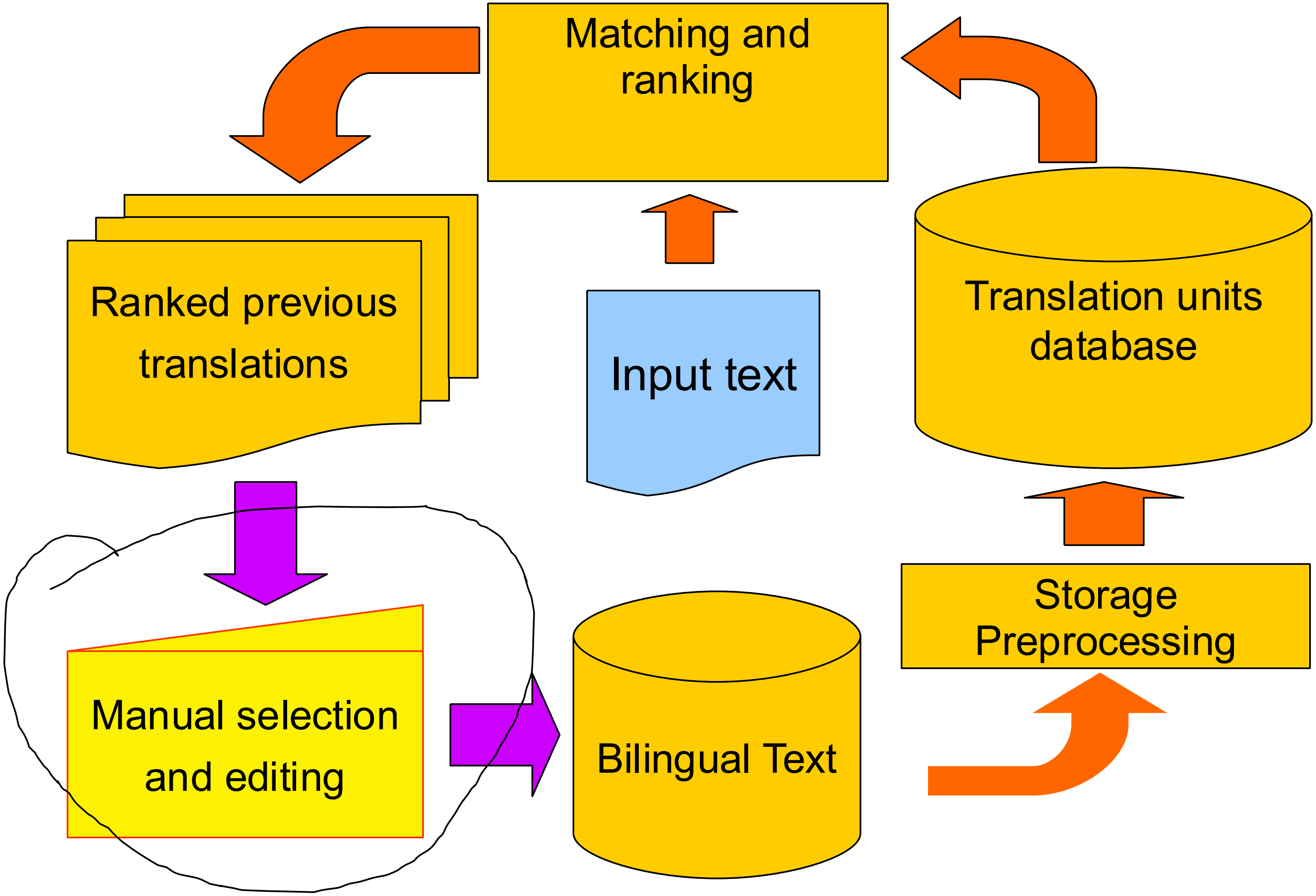
If a sentence has been translated already, use that translation as draft.

At about the same time, a new tool for translators was being discussed by developers. Like EBMT, it used a corpus of already-translated examples to serve as models for the new translation, but crucially, it was the human users, not the computer itself, who should determine exactly how to use the examples in producing a new translation. This tool is of course now widely known as a Translation Memory System (TMS).

Example-Based MT


Fragment Retrieval, Combination and Ranking
Translation
Input text
Translation units database
Storage Preprocessing
Bilingual Text

Translation Memory



Matching and ranking

Ranked previous translations

Input text

Translation units database

Manual selection and editing

Bilingual Text

Storage Preprocessing

Example-based Machine Translation

- Aim: to produce a correct translation
- Makes efficient use of previous translations
- Examples used for disambiguation

- Example selection and combination is difficult
- Context dependent phenomena such as anaphora difficult to handle
- Sally arrived but nobody saw her.

Translation Memory

- High quality translation when good matches are found
- Useful for highly repetitive or updated documents

- Does not combine translations from different sentences
- Not suitable for general/unrestricted text

Translation Memory Tools

- Translation memory tools are specifically designed to recycle previously created translations as much as possible.

- Translation memory is a tool that allows the user to store translated phrases or sentences in a special database for local re-use or shared use over a network.

- Translation memory systems work by matching terms and sentences in the database with those in the source text.
  - If a match is found, the system proposes the ready-made translation in the target language

Neural Machine Translation

- Translation memory tools are specifically designed to recycle previously created translations as much as possible.

- Translation memory is a tool that allows the user to store translated phrases or sentences in a special database for local re-use or shared use over a network.

- Translation memory systems work by matching terms and sentences in the database with those in the source text.
  - If a match is found, the system proposes the ready-made translation in the target language

Conclusion

- Computing is now multilingual and global
- Language aware software is important.
- Standards play a major role in simplifying multilingual software (UNICODE)
- Different paradigms have different strengths
  - Enough effort devoted to a paradigm or system will render it useful, regardless of its theoretical underpinnings
- Outstanding problems in MT: ambiguity and coverage

Further Reading

- Books:

Neural Network Methods in Natural Language Processing, 2017.

Syntax-based Statistical Machine Translation, 2017.

Deep Learning, 2016.

Statistical Machine Translation, 2010.

Handbook of Natural Language Processing and Machine Translation, 2011.

Artificial Intelligence, A Modern Approach, 3rd Edition, 2009.

Papers:

A Statistical Approach to Machine Translation, 1990.

Review Article: Example-based Machine Translation, 1999.

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014.

Neural Machine Translation by Jointly Learning to Align and Translate, 2014.

Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

Sequence to sequence learning with neural networks, 2014.

Recurrent Continuous Translation Models, 2013.

Continuous space translation models for phrase-based statistical machine translation, 2013.