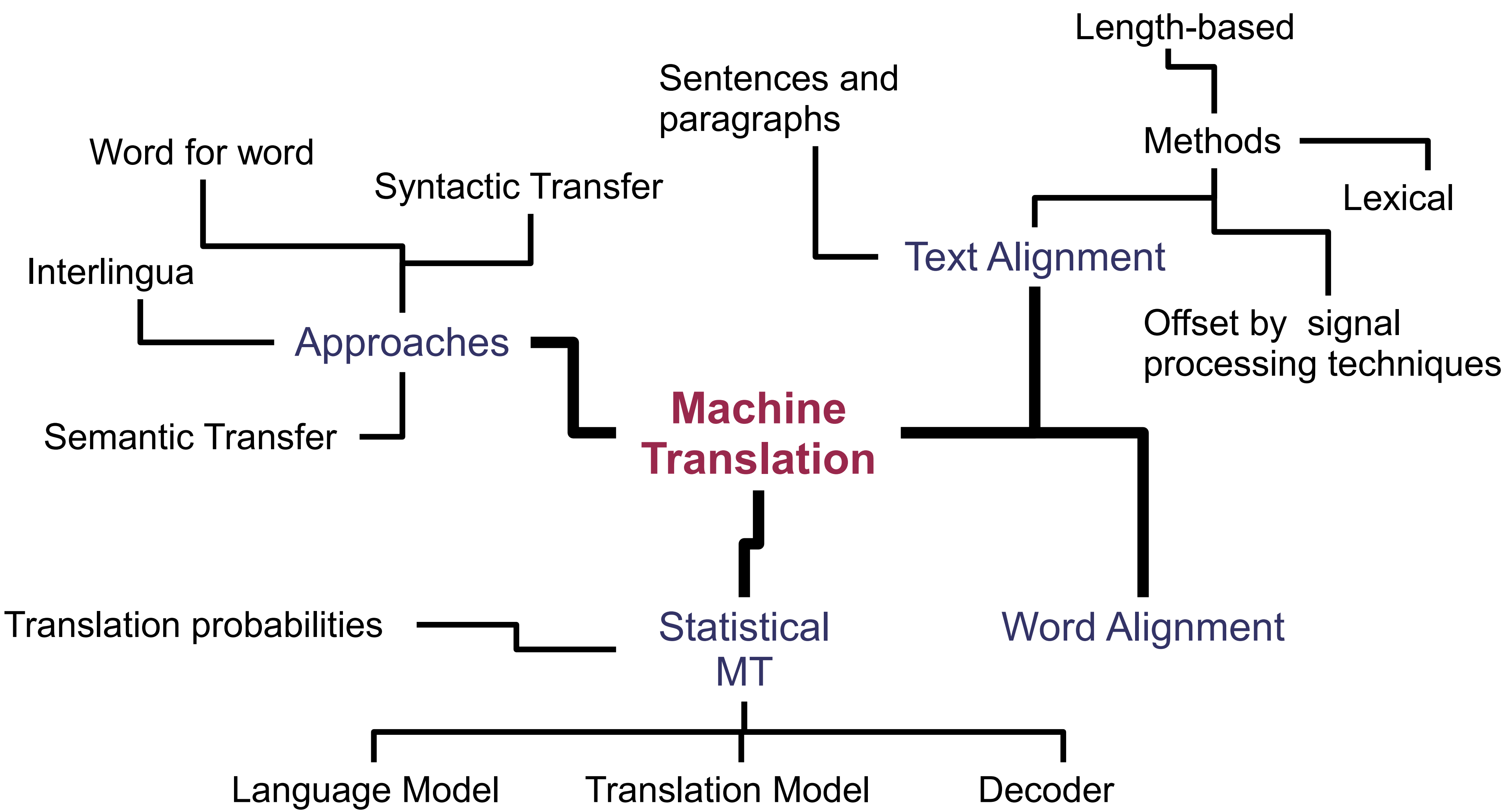
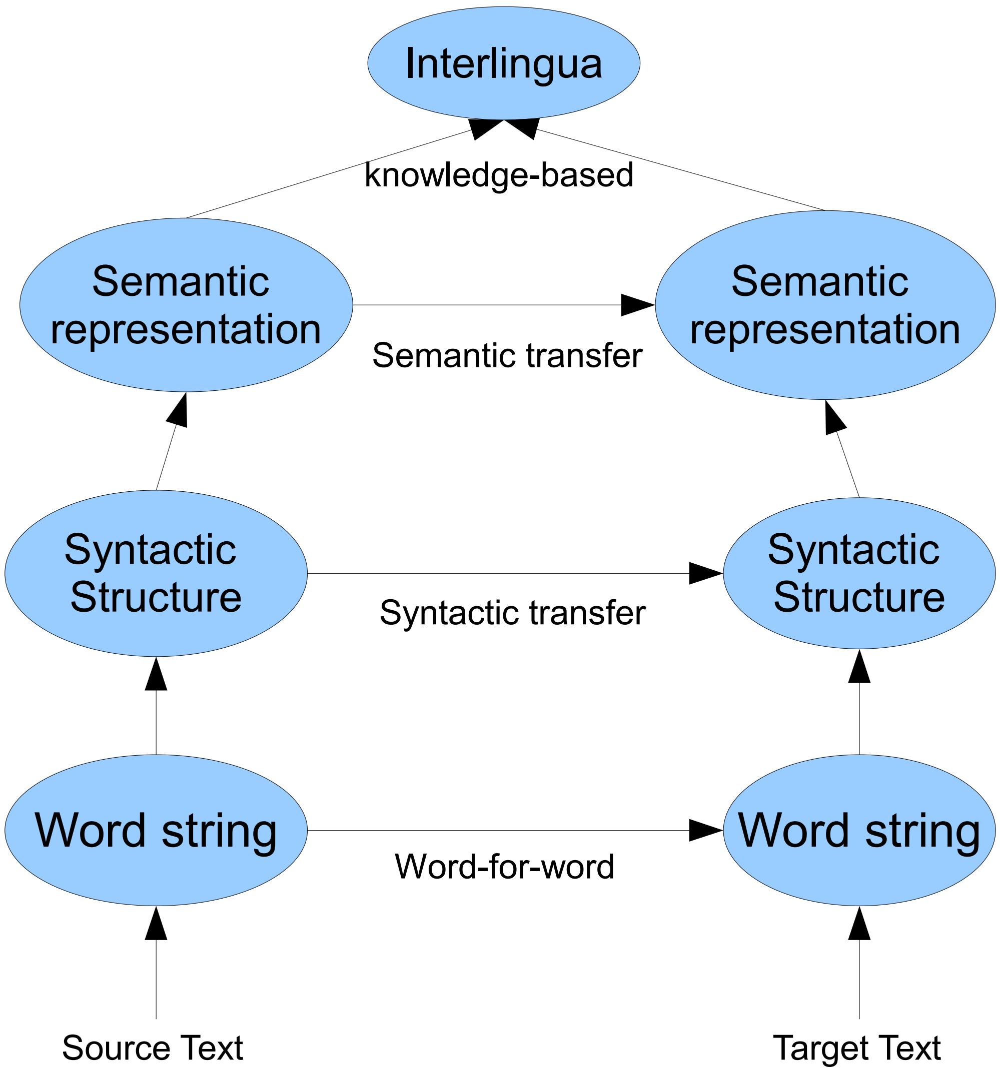
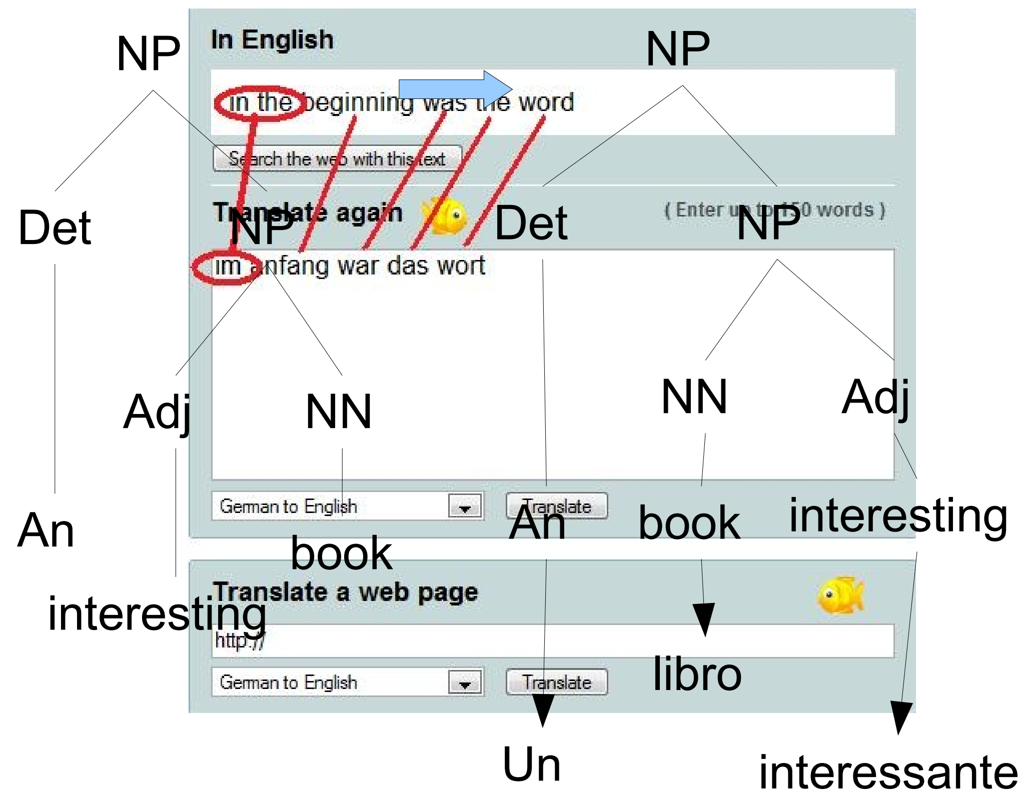
Statistical Alignment and Machine Translation



Approaches

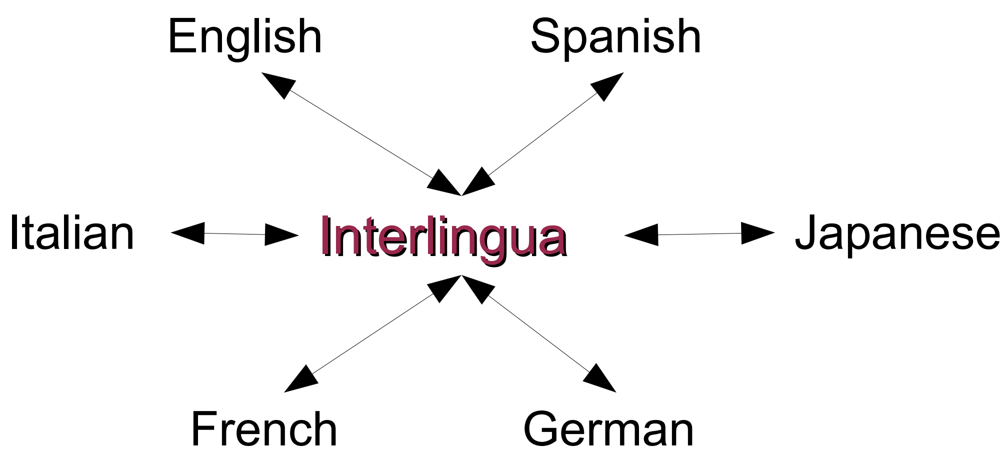


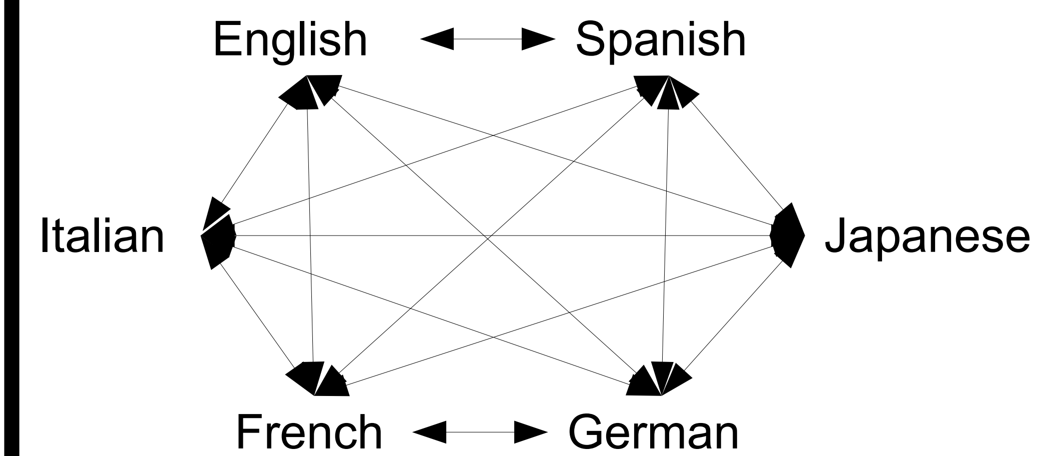
Parse the sentence ➔ Transform the sentence➔ Translate the words



Approaches: Interlingua

- Independent of the way particular languages express meaning
- Need only *O(n)* translation systems





Text Alignment

- Parallel texts (bitexts):
  - which paragraphs or sentences in one language correspond to which paragraphs or sentences in another language
  - Which words tend to be translated by which other words
- Aligning sentences and paragraphs
  - Some group of sentences in one language corresponds in context to some group of sentences in the other language
  - 1:1 one sentence in the source document corresponds to one sentence in the target document
  - n:m *n* sentences in the source document correspond to *m* sentences in the target document
  - How much content has to overlap between sentences?
  - Problem of crossing dependecies

Methods in Text Alignment

- *Length-Based Approaches:* compare the lengths of units of text in the parallel corpora (short sentences will be translated as short sentences and long sentences as long sentences)
- *Offset Alignment by Signal Processing Techniques*: attemp to align position offsets in the two parallel texts
- *Lexical Methods:* Use lexical information to align beads of sentences.

Length-based approaches

- Gale & Church (1993)
  - Find the alignment ***A*** with highest probability given the two parallel texts ***S*** and ***T***
  - To estimate the probabilities: decompose the aligned texts into a sequence of aligned beads
- Brown et al. (1991)
  - Similar to Gale & Church, but works by comparing sentence length in words (not characters)
- Wu(1994)
  - Unrelated languages (English and Cantonese)
  - Uses lexical cues

Offset Alignment by Signal Processing Techniques

- Church (1993)
  - Alignment using cognates (words that are similar across languages)

    supérieur (French) – superior (English)
  - Find cognates at the level of character sequences
  - Dot-plot construction
- Fung & McKeown(1994)
  - Seek an algorithm that will work without having found sentence boundaries, in only roughly parallel texts, and with unrelated language pairs
  - A small bilingual dictionary gives points of alignment
  - For each words, a signal is produced:

    an arrival vector of integer numbers giving the

    number of words between eachoccurence of the word at hand
  - A measure of similarity between signals is calculated using Dynamic Time Warping

Word offset: (1, 263, 267, 519)
Arrival vector: (262, 4, 252)

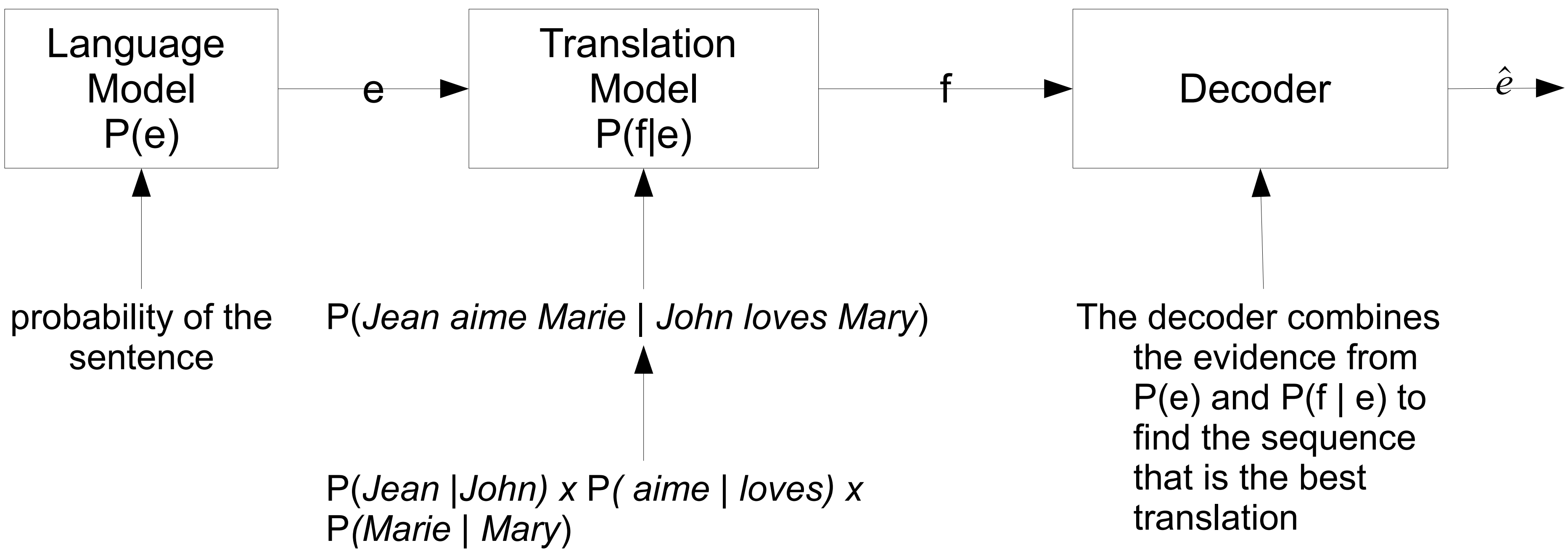
Lexical Methods

- Kay & Röscheisen(1993)
  - Assumption: two words should correspond if their distributions are the same
  - Steps:
    - Assume the first and last sentences of the texts align. These are the initial ***anchors***.
    - Then, until most sentences are aligned:
      - Form an ***envelope*** of possible alignments.
      - Choose pairs of words that tend to co-occur in these potential partial alignments.
      - Find pairs of source and target sentences which contain many possible lexical correspondences.
- Chen (1993)
  - Sentence alignment by constructing a simple word-to-word translation model
  - The best alignment is the one that maximizes the likelihood of generating the corpus given the translation model
- Haruno & Yamazaki (1996)
  - Work with structurally different languages
  - Do lexical matching on content words only (use POS taggers)
  - Use an online dictionary to find matching word pairs

Word Alignment

- Derivation of bilingual dictionaries and terminological databases:
  - text alignment is extended to a word alignment.
  - some criterion (frequency) is used to select aligned pairs
- Word alignment based on measures of association (χ^2)
  - works well unless one word in L1 frequently occurs with more than one word in L2. Then, it is useful to assume a one-to-one correspondence.

Statistical Machine Translation



$$\arg\max_e P(e \mid f) = \arg\max_e P(f \mid e) P(e)$$

Translation Probabilities

- Are estimated using the EM algorithm, which solves the credit assignment problem
- Random initialization of the translation probabiities
- Compute the expected number of times we will find w_f in the French sentence given that we have w_e in the English sentence
- Reestimate the translation probabilities from the expectations

Problems (Brown et al. 1990, 1993)

- Fertility is asymmetric
- Independence assumptions
- Sensitivity to training data
- Efficiency

Lack of linguistic knowledge

- No notion of phrases
- Non-local dependencies
- Morphology
- Sparse data problems