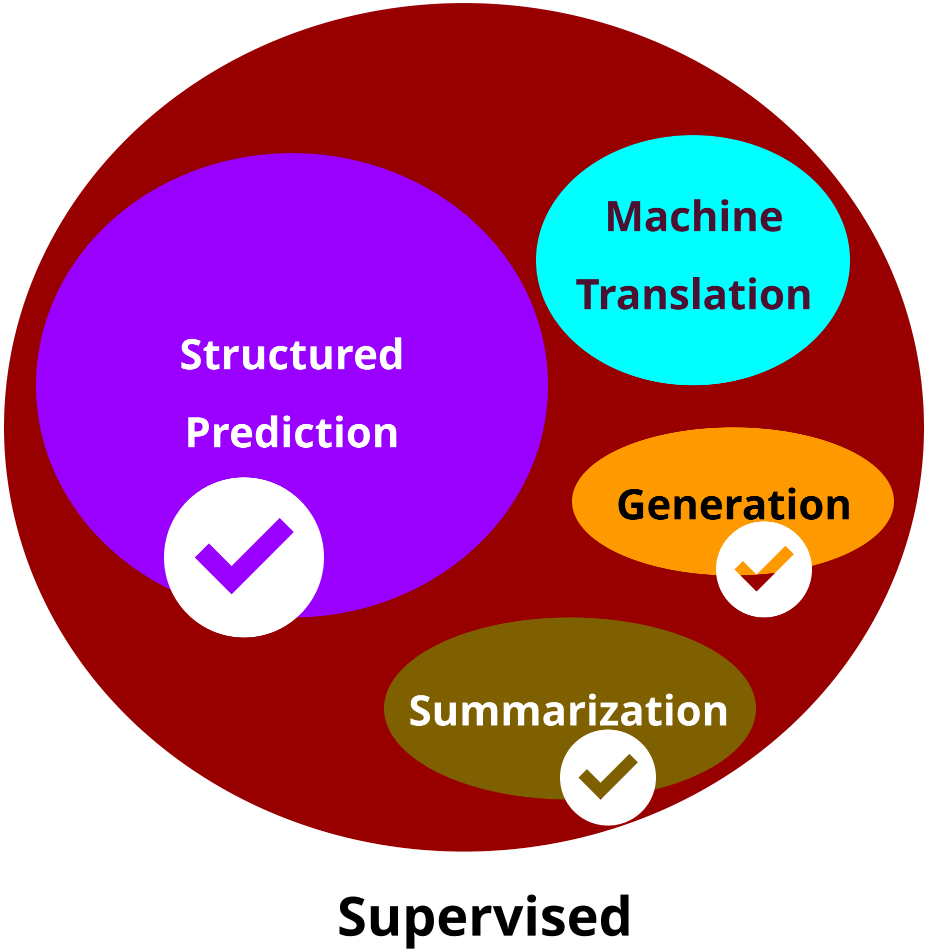
50.040
Natural Language Processing

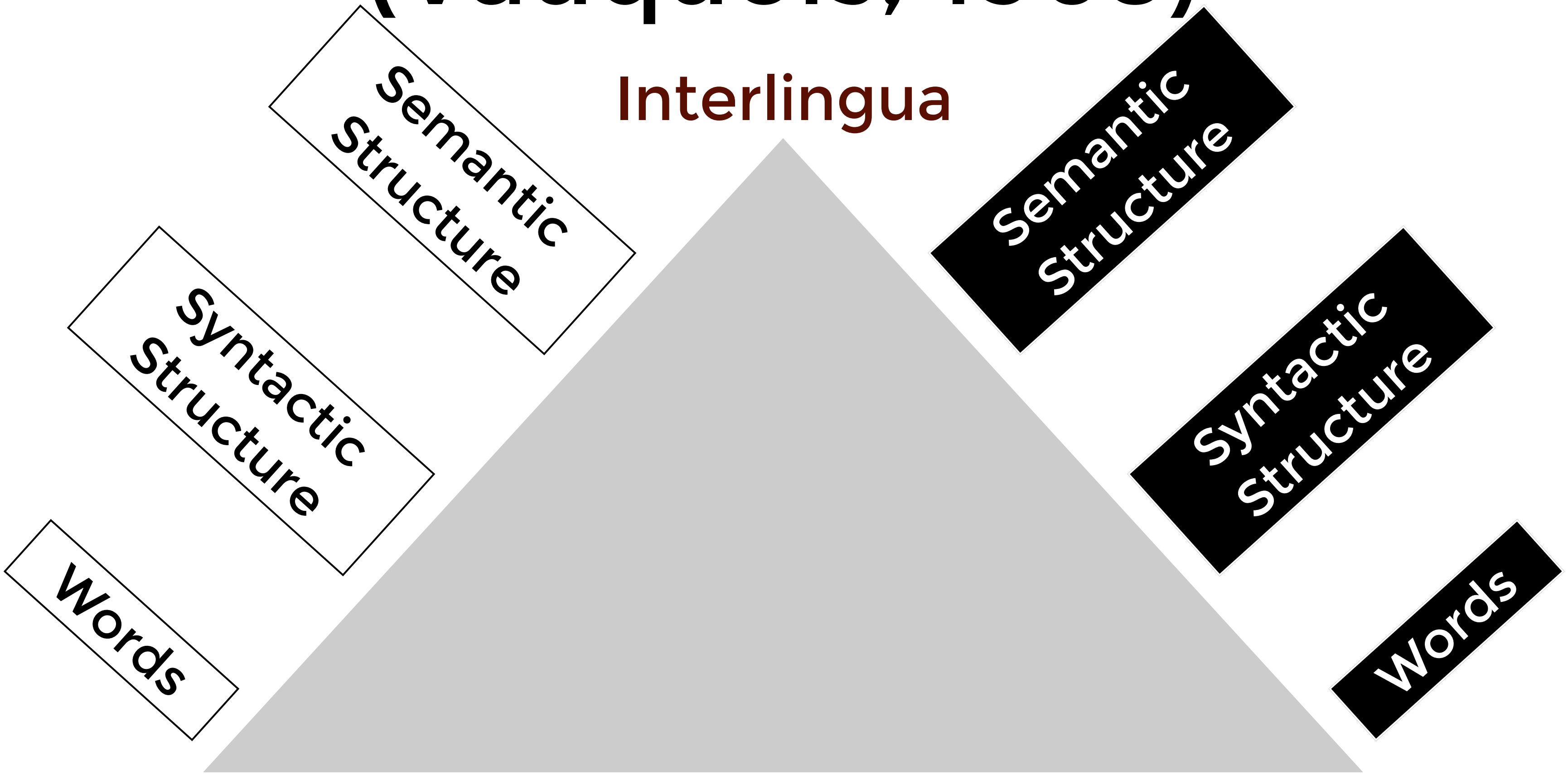
Lu, Wei

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

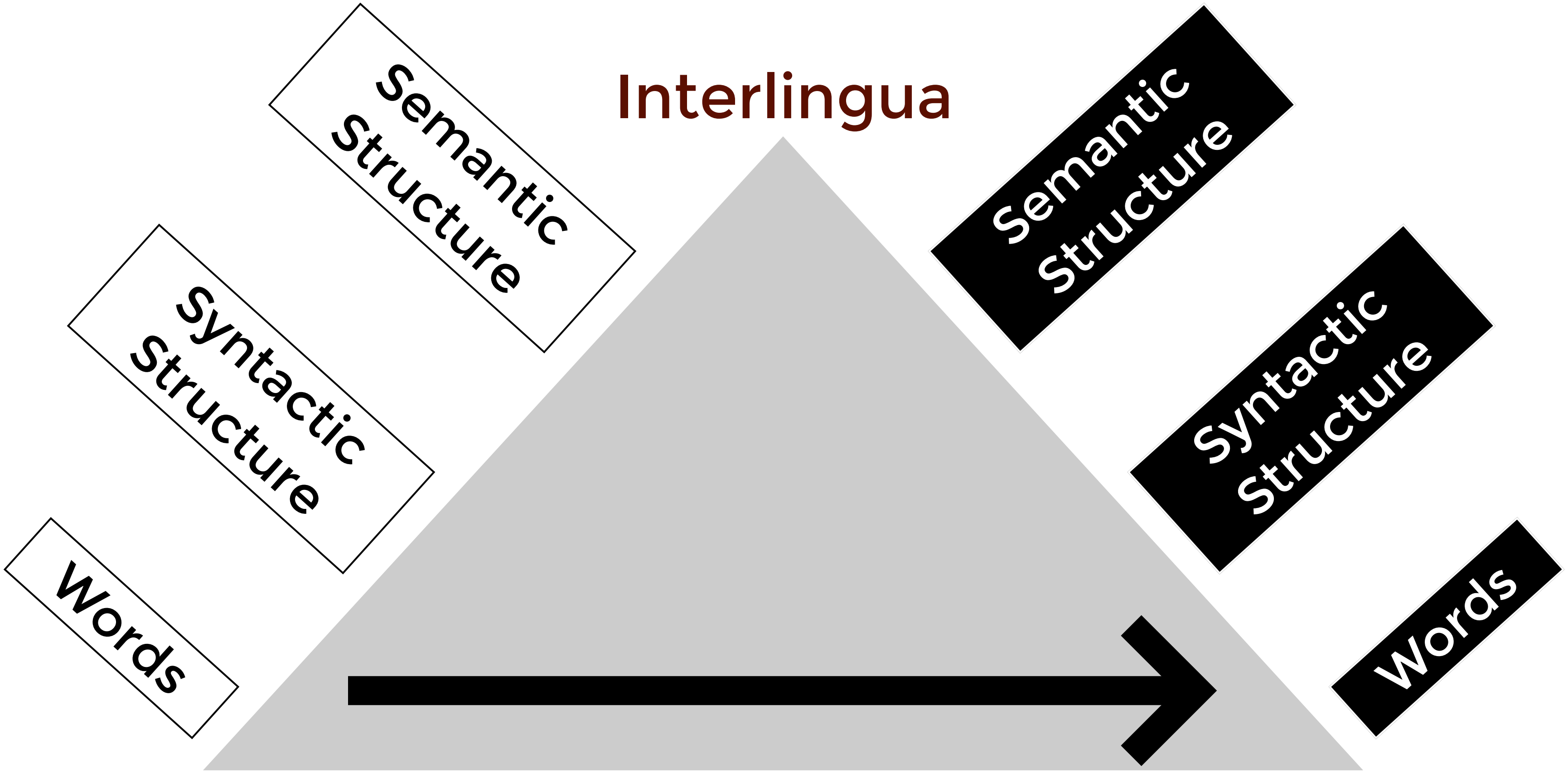
Tasks in NLP


Machine
Translation
Structured
Prediction
Generation
Summarization
Supervised

Machine Translation
(Vauquois, 1968)

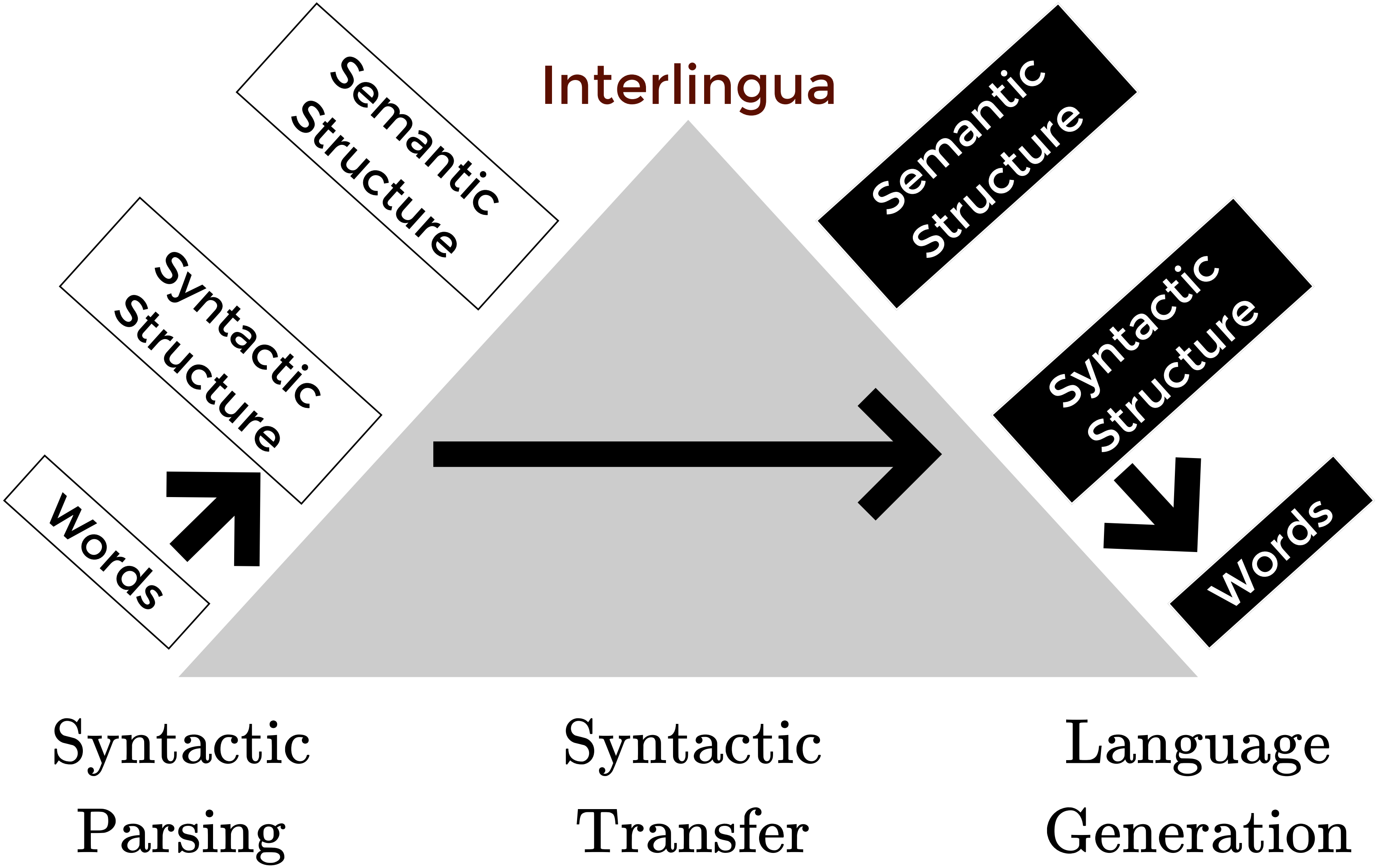


Machine Translation

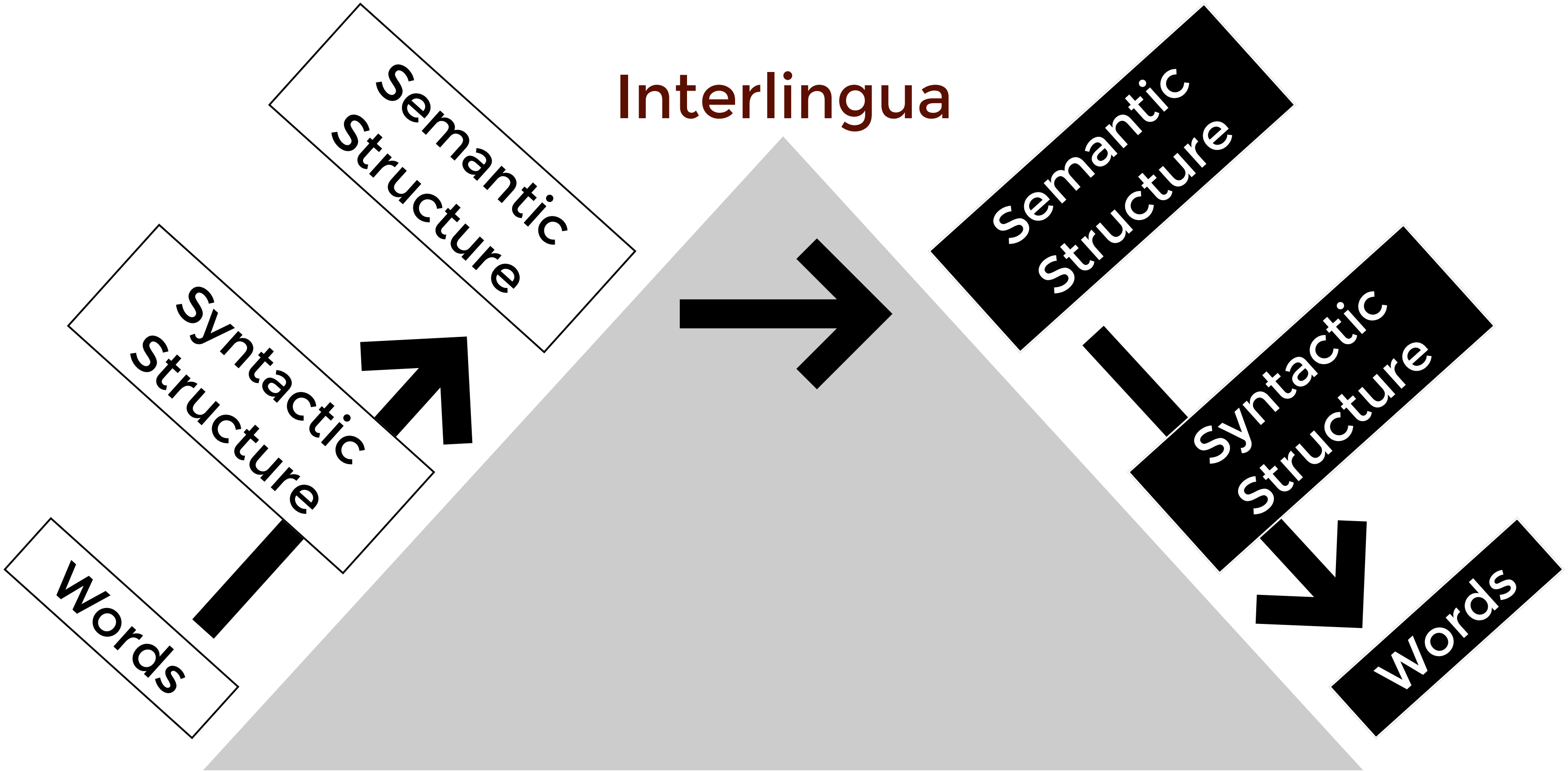


Text-to-text Problem

Machine Translation



Machine Translation

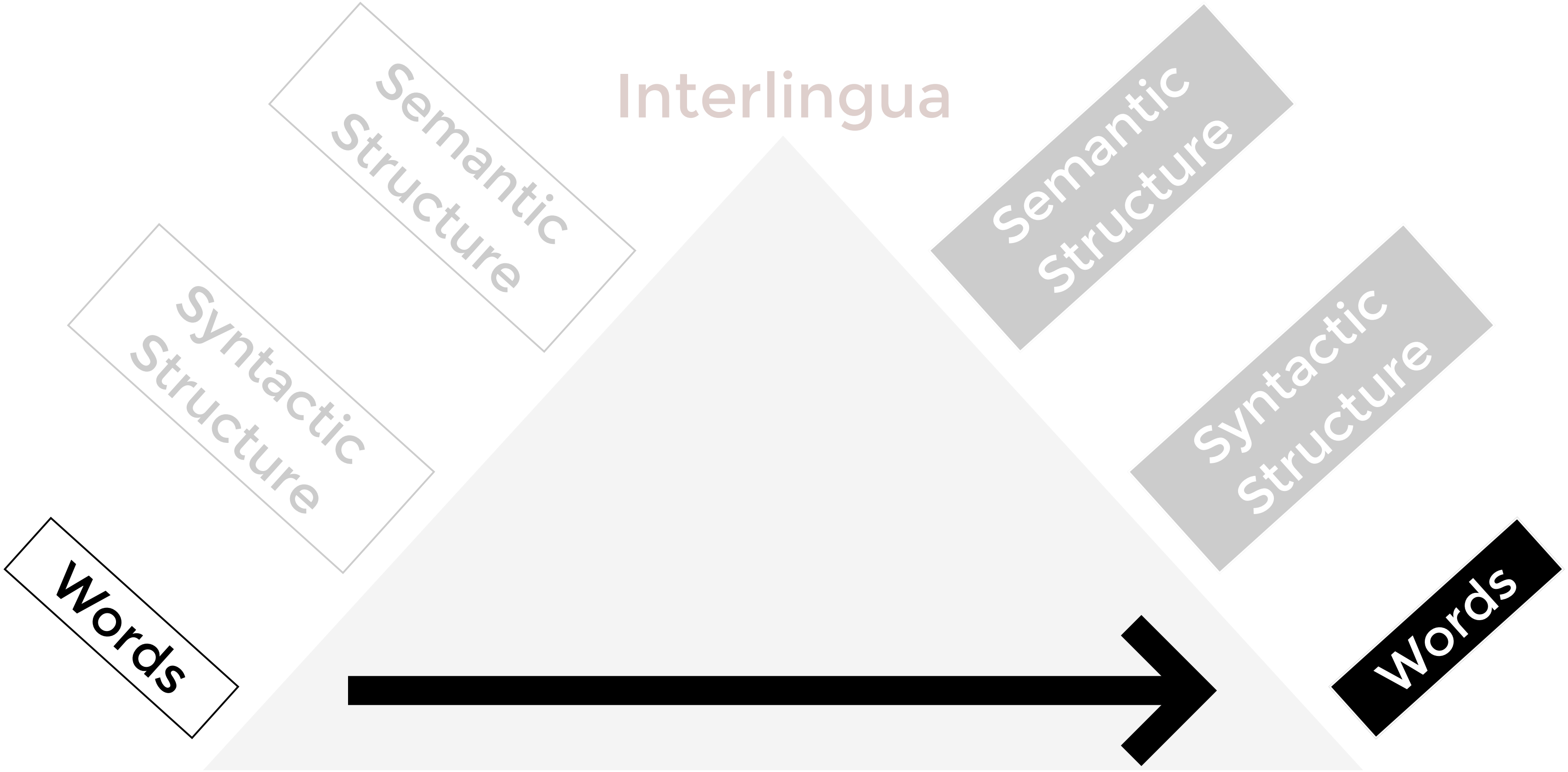


Semantic Parsing

Semantic Transfer

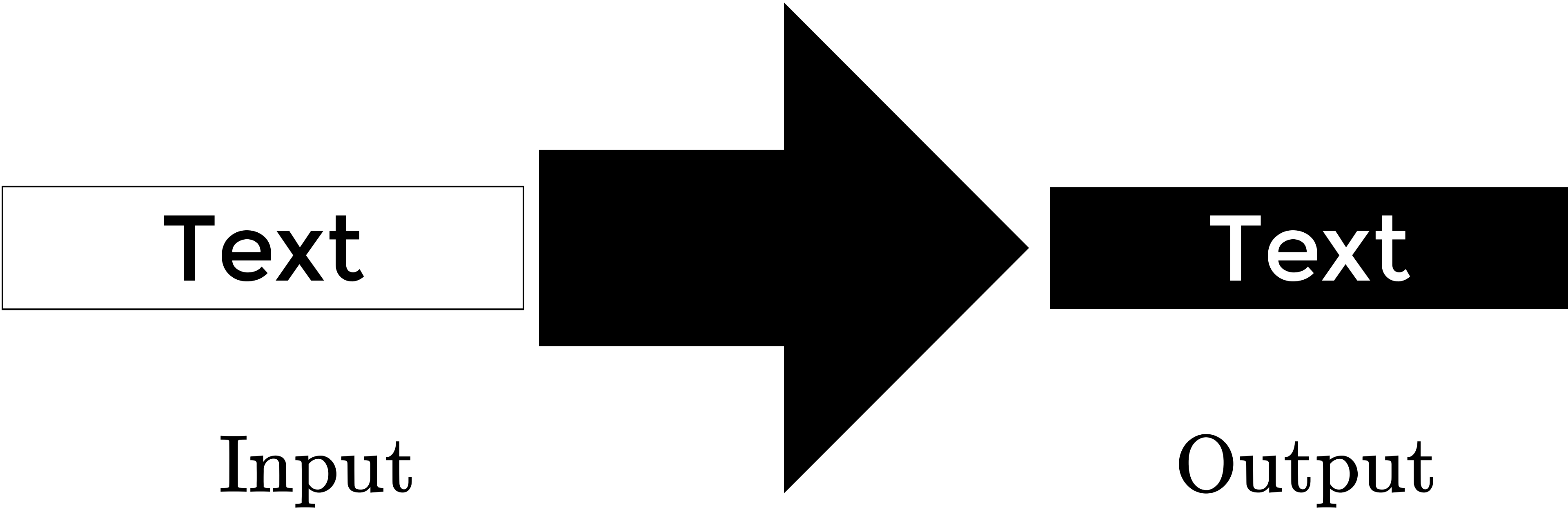
Language Generation

Machine Translation



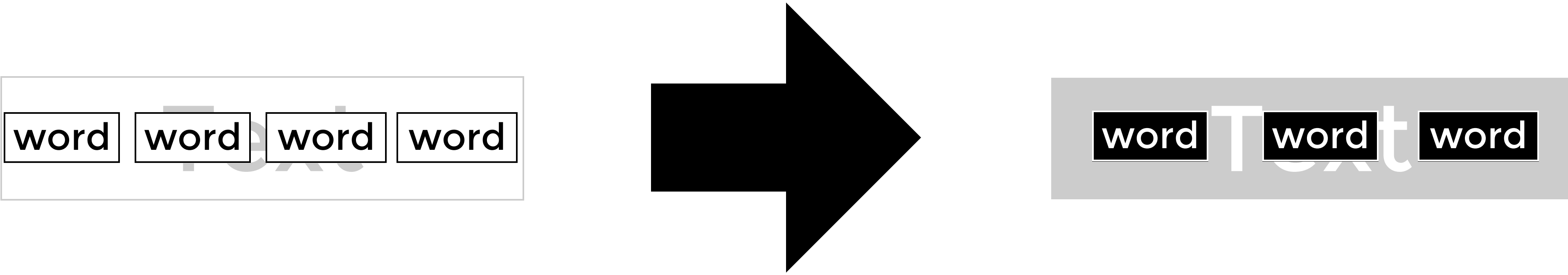
Text-to-text Problem

Machine Translation

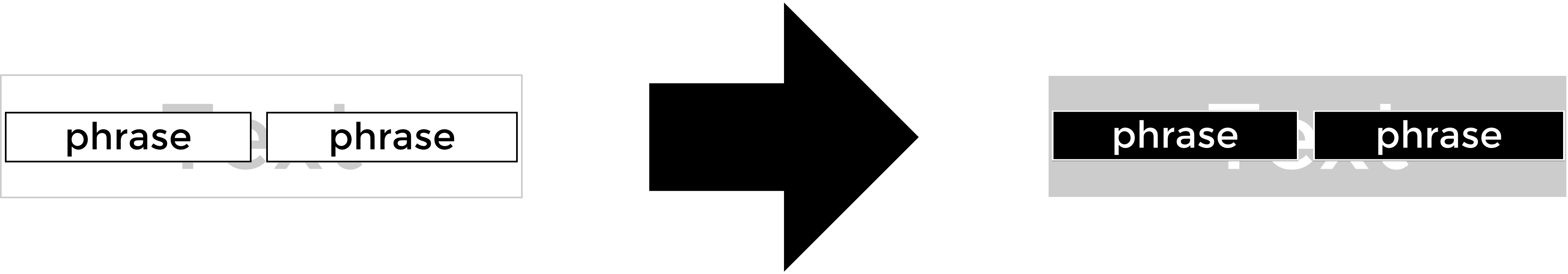


Text-to-text Problem

Machine Translation

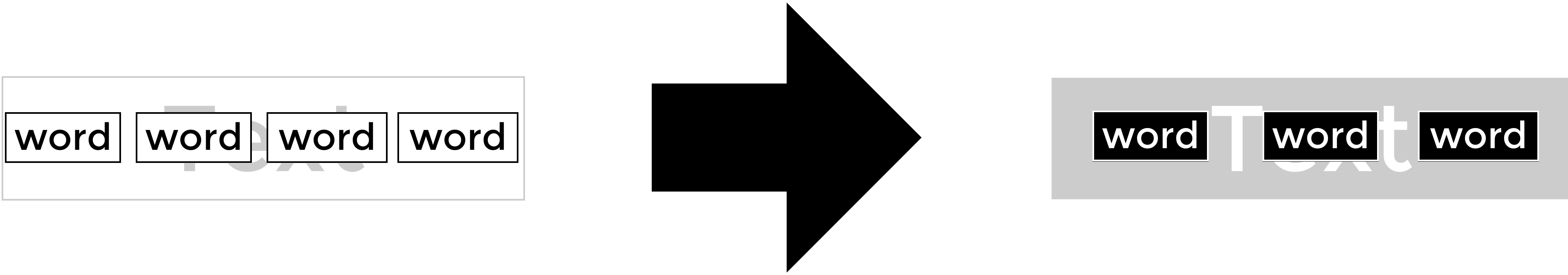


Word-based Translation



Phrase-based Translation

Machine Translation



Word-based Translation

Phrase-based Translation

Machine Translation

The convention is to assume we are translating from French (foreign language) into English.

English sentence

French sentence

$$\boldsymbol{e}^* = \arg\max_{\boldsymbol{e}} p(\boldsymbol{e} | \boldsymbol{f})$$

Noisy-Channel

The convention is to assume we are translating from French (foreign language) into English.

English sentence

French sentence

$$
\begin{aligned}
\boldsymbol{e}^* &= \arg\max_{\boldsymbol{e}} p(\boldsymbol{e}|\boldsymbol{f}) \\
&= \arg\max_{\boldsymbol{e}} p(\boldsymbol{f}|\boldsymbol{e})p(\boldsymbol{e})/p(\boldsymbol{f}) \\
&= \arg\max_{\boldsymbol{e}} p(\boldsymbol{f}|\boldsymbol{e})p(\boldsymbol{e})
\end{aligned}
$$

Noisy-Channel

English $\rightarrow$ Channel $\rightarrow$ French

$$p(\boldsymbol{e})\,p(\boldsymbol{f}\,|\,\boldsymbol{e})$$

Noisy-Channel

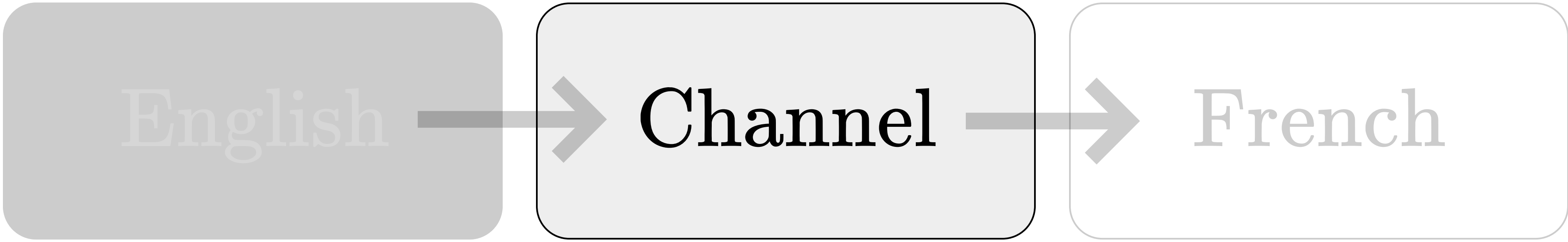
English → Channel → French

$$p(\boldsymbol{e})\,p(\boldsymbol{f}\,|\,\boldsymbol{e})$$

How good is the target sentence

Language Model

Noisy-Channel



$$p(\boldsymbol{e})\,p(\boldsymbol{f}|\boldsymbol{e})$$

How likely can we recover the source with the target sentence

Translation Model

IBM Models

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

We describe a series of five statistical models of the translation process and give algo estimating the parameters of these models given a set of pairs of sentences that are tran of one another. We define a concept of word-by-word alignment between such pairs of sen For any given pair of such sentences each of our models assigns a probability to each possible word-by-word alignments. We give an algorithm for seeking the most probable of t alignments. Although the algorithm is suboptimal, the alignment thus obtained accounts well the word-by-word relationships in the pair of sentences. We have a great deal of data in Frenc and English from the proceedings of the Canadian Parliament. Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.

1. Introduction

The growing availability of bilingual, machine-readable texts has stimulated interest in methods for extracting linguistically valuable information from such texts. For example, a number of recent papers deal with the problem of automatically obtaining pairs of aligned sentences from parallel corpora (Warwick and Russell 1990; Brown, Lai, and Mercer 1991; Gale and Church 1991b; Kay 1991). Brown et al. (1990) assert, and Brown, Lai, and Mercer (1991) and Gale and Church (1991b) both show, that it is possible to obtain such aligned pairs of sentences without inspecting the words that the sentences contain. Brown, Lai, and Mercer base their algorithm on the number of words that the sentences contain, while Gale and Church base a similar algorithm on the number of characters that the sentences contain. The lesson to be learned from these two efforts is that simple, statistical methods can be surprisingly successful in achieving linguistically interesting goals. Here, we address a natural extension of that work: matching up the words within pairs of aligned sentences.

In recent papers, Brown et al. (1988, 1990) propose a statistical approach to machine translation from French to English. In the latter of these papers, they sketch an algorithm for estimating the probability that an English word will be translated into any particular French word and show that such probabilities, once estimated, can be used together with a statistical model of the translation process to align the words in an English sentence with the words in its French translation (see their Figure 3).

* IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Translation Model

$$p(\boldsymbol{f}|\boldsymbol{e})$$

$$p(f_1, f_2, \ldots, f_m|e_1, e_2, \ldots, e_n)$$

$$\boldsymbol{p(m|n)}\, p(f_1, f_2, \ldots, f_m|e_1, e_2, \ldots, e_n, m)$$

The probability of the French sentence length, given the English sentence length

Translation Model

$$p(\boldsymbol{f}|\boldsymbol{e})$$

$$p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n)$$

$$p(m|n)\, p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n, m)$$

The probability of generating the French sentence, given the English and the expected length of French sentence

Translation Model

$$p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n, m)$$

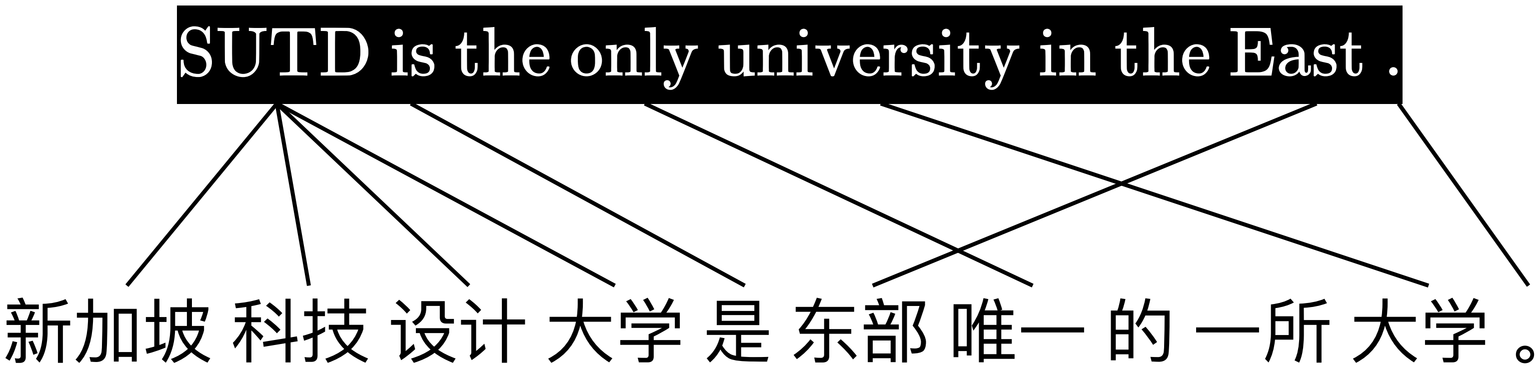
SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

How to generate each foreign word from the sequence of English words? Seems something is missing...

Translation Model

$$p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n, m)$$



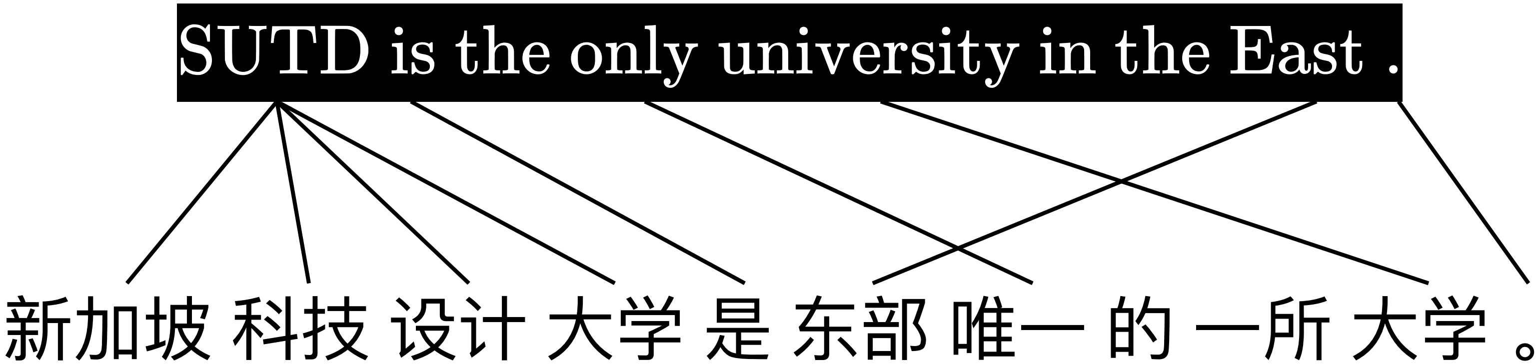
One possible alignment

Translation Model

$$p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_1, e_2, \ldots, e_n, m)$$

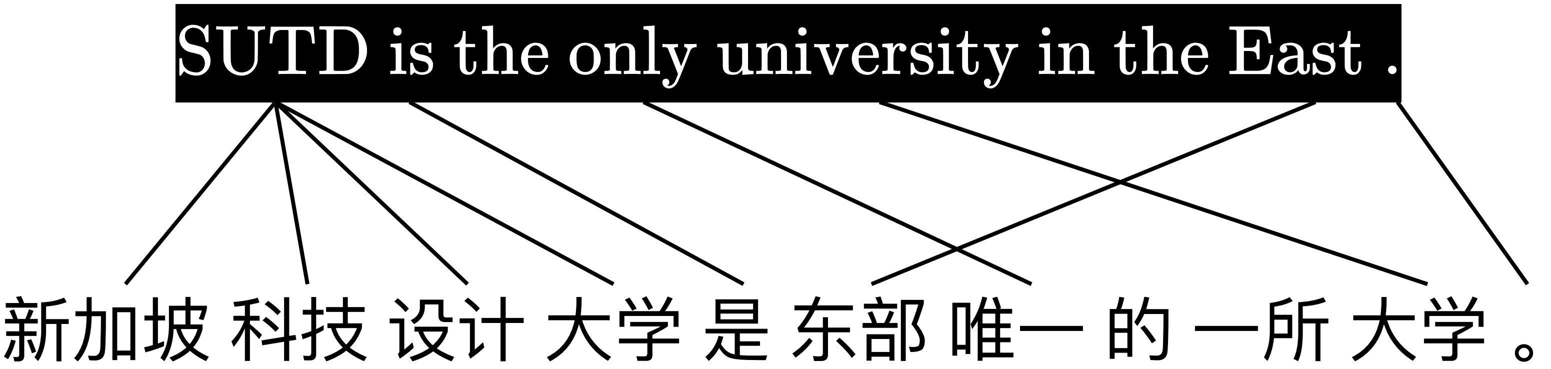
The position of the English word that should be aligned to the 1st French word



Translation Model

$$p(f_1, f_2, \ldots, f_m | e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_1, e_2, \ldots, e_n, m)$$

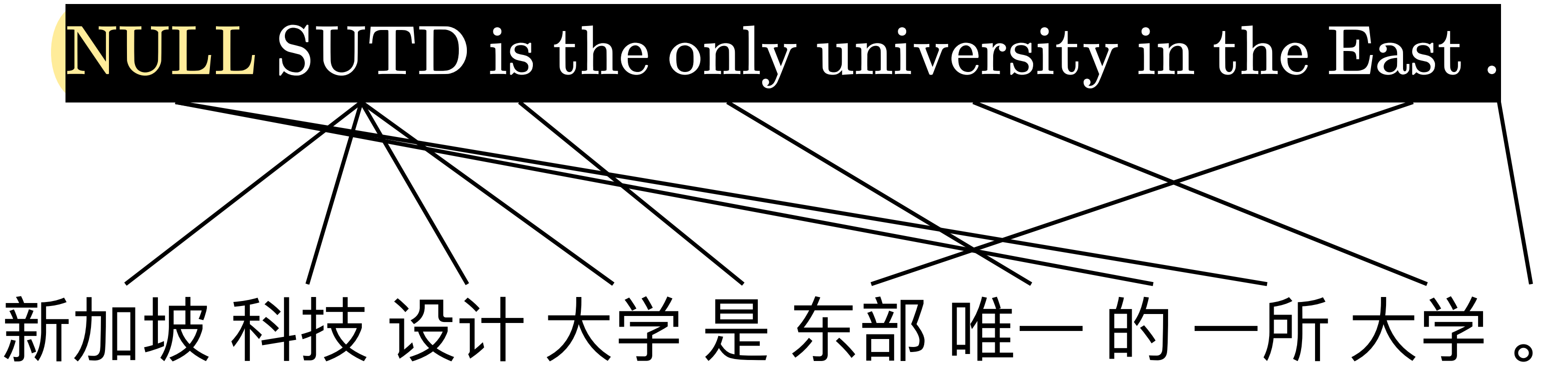


Can you figure out each alignment variable for this particular example above based on the alignment given? Do you see any issue when working on this?

Translation Model

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$



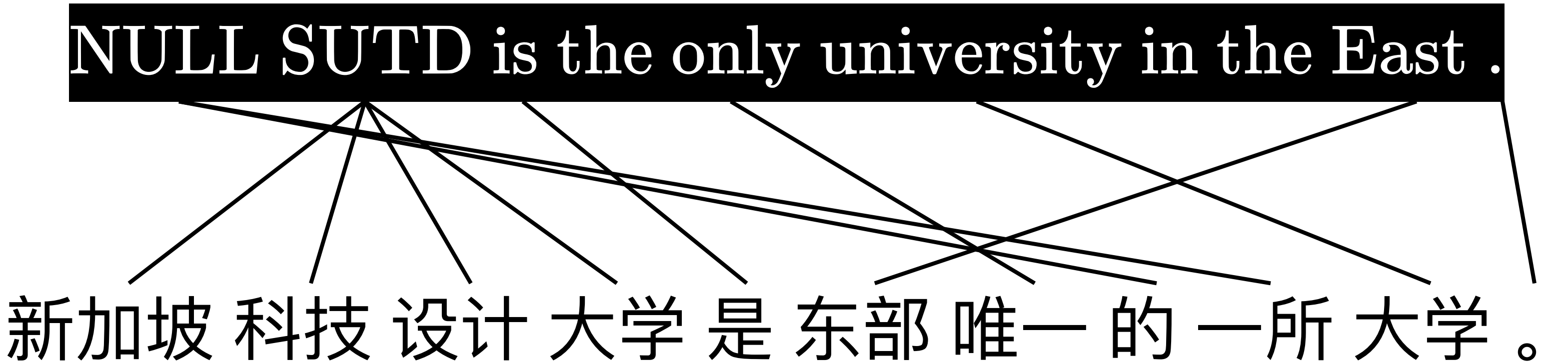
We need to introduce the special NULL tokens!

Translation Model

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\prod_{i=1}^{m} q(a_i | i, n, m) t(f_i | e_{a_i})$$



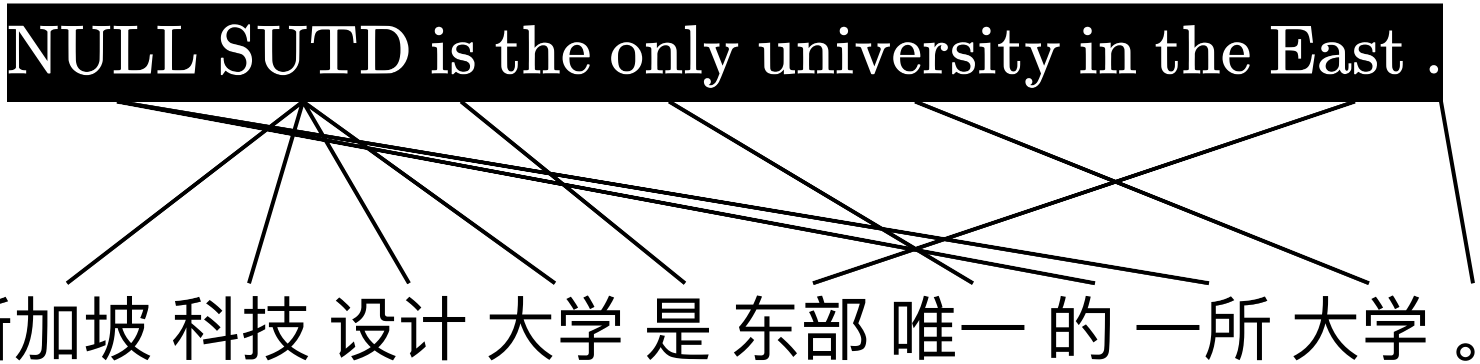
IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\prod_{i=1}^{m} q(a_i | i, n, m) t(f_i | e_{a_i})$$

$$\prod_{i=1}^{m} \frac{1}{n+1} t(f_i | e_{a_i})$$



IBM Model 1

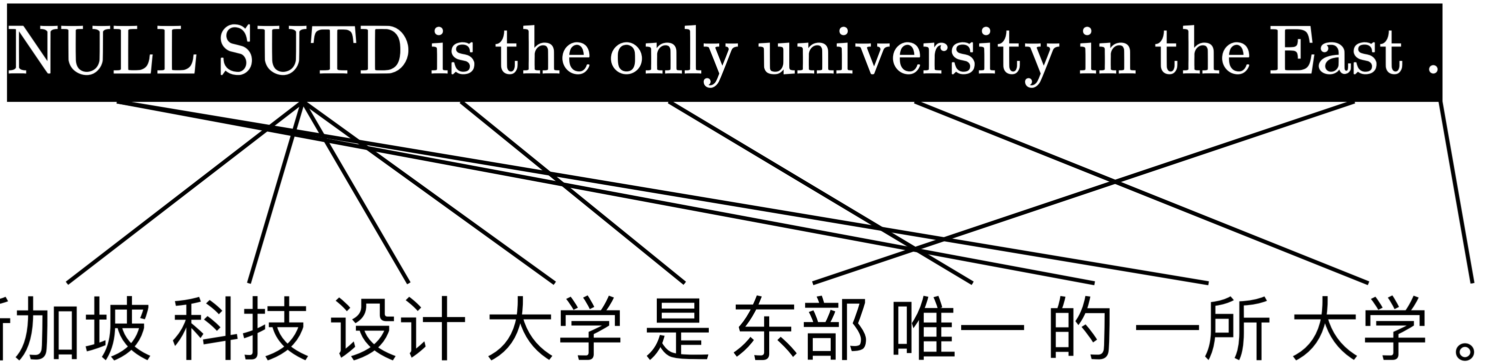
$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\prod_{i=1}^{m} q(a_i | i, n, m) t(f_i | e_{a_i})$$

$$\prod_{i=1}^{m} \frac{1}{n+1} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$



IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\sum_{a_1} \cdots \sum_{a_m} \frac{1}{(n+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\sum_{a_1} \cdots \sum_{a_m} \frac{1}{(n+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \sum_{a_1} \cdots \sum_{a_m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\sum_{a_1} \cdots \sum_{a_m} \frac{1}{(n+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \sum_{a_1} \cdots \sum_{a_m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \sum_{a_1} t(f_1 | e_{a_1}) \cdots \sum_{a_m} t(f_m | e_{a_m})$$

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\sum_{a_1} \cdots \sum_{a_m} \frac{1}{(n+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \sum_{a_1} \cdots \sum_{a_m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

$$\frac{1}{(n+1)^m} \sum_{a_1} t(f_1 | e_{a_1}) \cdots \sum_{a_m} t(f_m | e_{a_m})$$

$$\frac{1}{(n+1)^m} \left(\sum_{a_1} t(f_1 | e_{a_1}) \right) \ldots \left(\sum_{a_m} t(f_m | e_{a_m}) \right)$$

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\frac{1}{(n+1)^m} \Big(\sum_{a_1} t(f_1 | e_{a_1}) \Big) \ldots \Big(\sum_{a_m} t(f_m | e_{a_m}) \Big)$$

$$\frac{1}{(n+1)^m} \prod_{i=1}^{m} \Big(\sum_{j=0}^{n} t(f_i | e_j) \Big)$$

Instance-level objective (log-likelihood):

$$-m \log(n+1) + \sum_{i=1}^{m} \log \Big(\sum_{j=0}^{n} t(f_i | e_j) \Big)$$

A constant term

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

We would like to maximize:

$$\sum_{i=1}^{m} \log \left(\sum_{j=0}^{n} t(f_i | e_j) \right)$$

Expectation-Maximization

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$\sum_{i=1}^{m} \log \left(\sum_{j=0}^{n} t(f_i | e_j) \right)$$

Initialization

randomly initialize the t probabilities

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

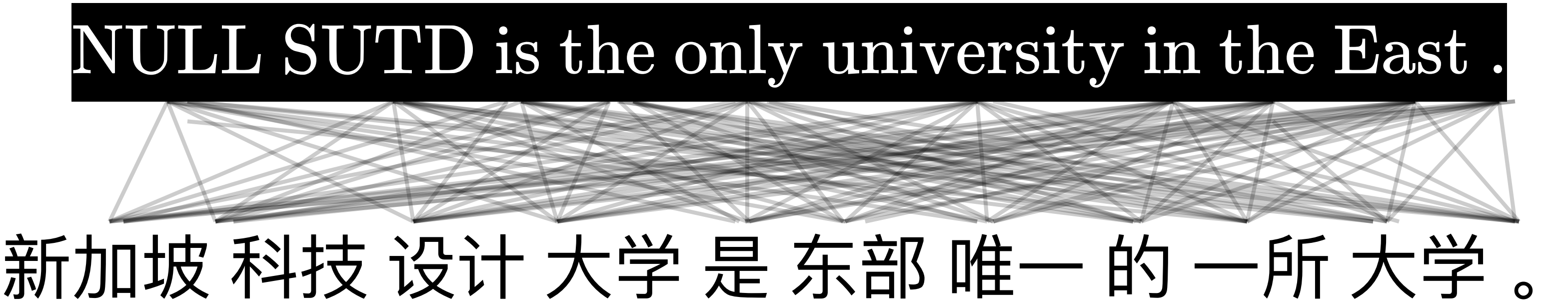
IBM Model 1

$$\sum_{i=1}^{m} \log\left(\sum_{j=0}^{n} t(f_i|e_j)\right)$$

Expectation

find for each f_i its membership

soft or hard alignment with English words



IBM Model 1

Soft EM

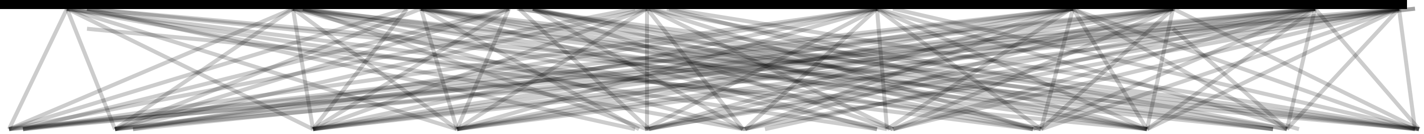
$$\sum_{i=1}^{m} \log \Big(\sum_{j=0}^{n} t(f_i|e_j) \Big)$$

Expectation

$$\text{count}^{(k)}(i,j) = \frac{t(f_i|e_j)}{\sum_{j'=0}^{n} t(f_i|e_{j'})}$$

In the k-th instance, how likely is the target word i aligned with the source word j (i.e., the expected number of times the target word i is aligned with the source word j).

NULL SUTD is the only university in the East .



新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

Hard EM

$$\sum_{i=1}^{m} \log \left(\sum_{j=0}^{n} t(f_i | e_j) \right)$$

Expectation

$$\text{count}^{(k)}(i,j) = \begin{cases} 1 & j = \arg\max_{j'} t(f_i | e_{j'}) \\ 0 & \text{o.w.} \end{cases}$$

In the k-th instance, what is the most probable source word j that the target word i should be aligned to?

NULL SUTD is the only university in the East .

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$\sum_{i=1}^{m} \log\left(\sum_{j=0}^{n} t(f_i|e_j)\right)$$

Maximization

update the model parameters t

NULL SUTD is the only university in the East .

Update model parameters

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 1

$$\sum_{i=1}^{m} \log \left(\sum_{j=0}^{n} t(f_i | e_j) \right)$$

Maximization

$$t(f|e) = \frac{\text{count}(e,f)}{\text{count}(e)}$$

$$\text{where: count}(e, f) = \sum_{k, f_i = f, e_j = e} \text{count}^{(k)}(i, j)$$

$$\text{count}(e) = \sum_{k, i, e_j = e} \text{count}^{(k)}(i, j)$$

NULL SUTD is the only university in the East .

Update model parameters

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 2

$$p(f_1, f_2, \ldots, f_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$p(f_1, f_2, \ldots, f_m, a_1, a_2, \ldots, a_m | e_0, e_1, e_2, \ldots, e_n, m)$$

$$\prod_{i=1}^{m} q(a_i | i, n, m) t(f_i | e_{a_i})$$

No longer a uniform distribution!
It models absolute reordering!

IBM Model 2

We need to estimate the additional q parameters using EM!

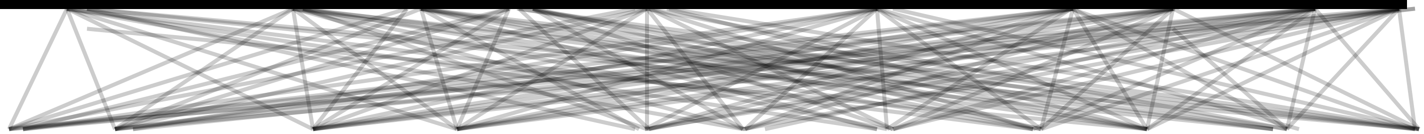
Expectation

still,

find for each f_i its membership

(soft/hard alignment with English words)

NULL SUTD is the only university in the East .



新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 2

We need to estimate the additional q parameters using EM!

Expectation

$$\text{count}^{(k)}(i, j, n, m) = \frac{q(j|i,n,m)t(f_i|e_j)}{\sum_{j'=0}^{n} q(j'|i,n,m)t(f_i|e_{j'})}$$

(soft/hard alignment with English words)

In the k-th instance (English and French sentence lengths are n an m respectively), the expected number of times we see the i-th French word is aligned to the j-th English word.

IBM Model 2

We need to estimate the additional q parameters using EM!

Maximization

additionally,

update the model parameters q

NULL SUTD is the only university in the East .

Update model parameters

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

IBM Model 2

We need to estimate the additional q parameters using EM!

Maximization

$$q(j|i,n,m) = \frac{\text{count}(i,j,n,m)}{\text{count}(i,n,m)}$$

$$\text{where: count}(i,j,n,m) = \textstyle\sum_k \text{count}^{(k)}(i,j,n,m)$$

$$\text{count}(i,n,m) = \textstyle\sum_{k,j} \text{count}^{(k)}(i,j,n,m)$$

IBM Model 2 captures absolute reordering on top of IBM model 1. Can we do something better?

IBM Models

Model 1	Lexical translation
Model 2	Adds absolute ordering
Model 3	Adds fertility
Model 4	Relative reordering
Model 5	Fixes deficiency

Optional

IBM Models

A Systematic Comparison of Various Statistical Alignment Models

Franz Josef Och*
University of Southern California

Hermann Ney†
RWTH Aachen

We present and compare various methods for computing word alignments using sta heuristic models. We consider the five alignment models presented in Brown, Della Piet Pietra, and Mercer (1993), the hidden Markov alignment model, smoothing techniqu refinements. These statistical models are compared with two heuristic models based on the coefficient. We present different methods for combining word alignments to perform a symmet tion of directed statistical alignment models. As evaluation criterion, we use the quality of resulting Viterbi alignment compared to a manually produced reference alignment. We evalua the models on the German-English Verbmobil task and the French-English Hansards task. We perform a detailed analysis of various design decisions of our statistical alignment system and evaluate these on training corpora of various sizes. An important result is that refined alignment models with a first-order dependence and a fertility model yield significantly better results than simple heuristic models. In the Appendix, we present an efficient training algorithm for the alignment models presented.

1. Introduction

We address in this article the problem of finding the word alignment of a bilingual sentence-aligned corpus by using language-independent statistical methods. There is a vast literature on this topic, and many different systems have been suggested to solve this problem. Our work follows and extends the methods introduced by Brown, Della Pietra, Della Pietra, and Mercer (1993) by using refined statistical models for the translation process. The basic idea of this approach is to develop a model of the translation process with the word alignment as a hidden variable of this process, to apply statistical estimation theory to compute the "optimal" model parameters, and to perform alignment search to compute the best word alignment.

So far, refined statistical alignment models have in general been rarely used. One reason for this is the high complexity of these models, which makes them difficult to understand, implement, and tune. Instead, heuristic models are usually used. In heuristic models, the word alignments are computed by analyzing some association score metric of a link between a source language word and a target language word. These models are relatively easy to implement.

In this article, we focus on consistent statistical alignment models suggested in the literature, but we also describe a heuristic association metric. By providing a detailed description and a systematic evaluation of these alignment models, we give the reader various criteria for deciding which model to use for a given task.

* Information Science Institute (USC/ISI), 4029 Via Marina, Suite 1001, Marina del Rey, CA 90292.
† Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen–University of Technology, D-52056 Aachen, Germany.

GIZA++: The implementations of all IBM models

Question

How to make use of the models to translate a new sentence?

Noisy-Channel

English → Channel → French

$$p(\boldsymbol{e})\,p(\boldsymbol{f}\,|\,\boldsymbol{e})$$

Still a computationally difficult problem to find $\boldsymbol{e}$!

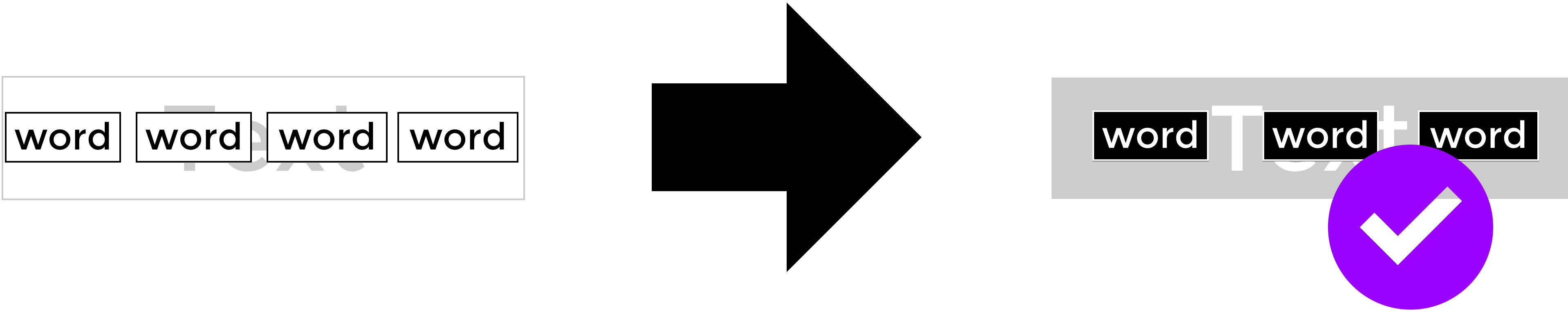
IBM Models

Pioneering work on word-level translation

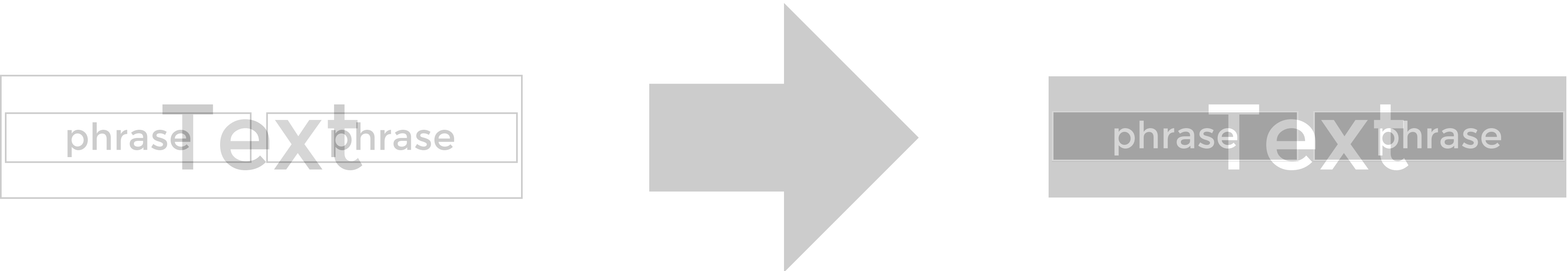
Not particularly useful models for translation themselves, but can yield useful alignment information for the training set

A first step towards building other advanced models such as phrase-based and syntax-based models

Machine Translation



Word-based Translation



Phrase-based Translation

Machine Translation

Phrase-based Translation