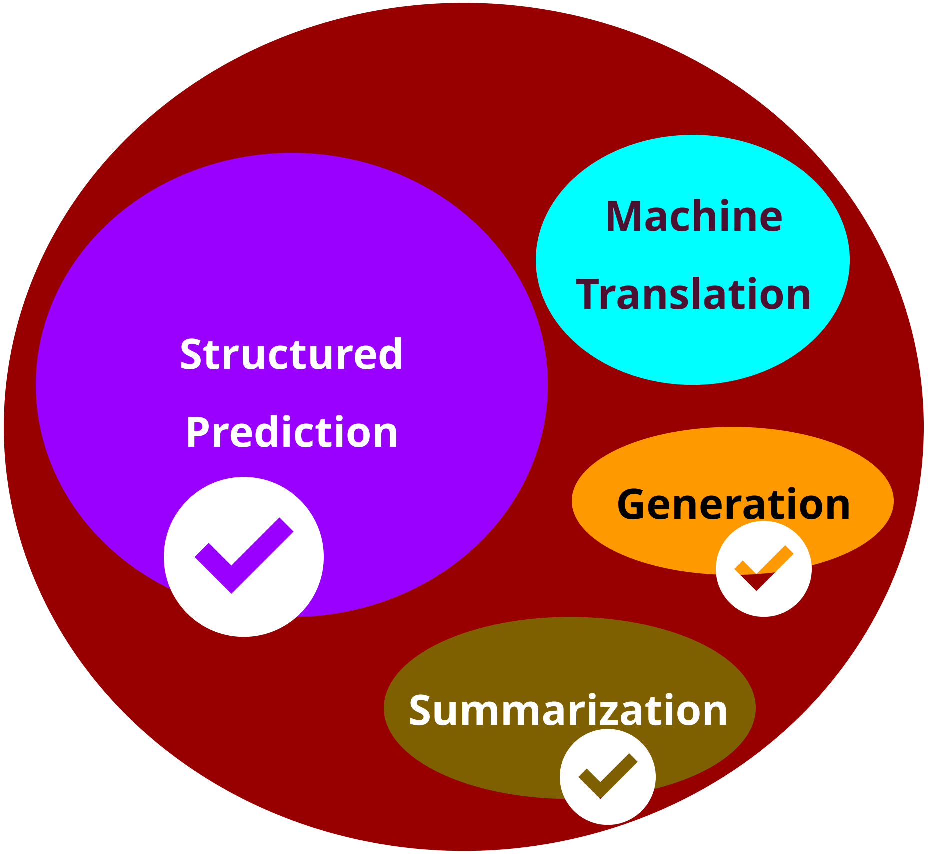
50.040
Natural Language Processing

Lu, Wei

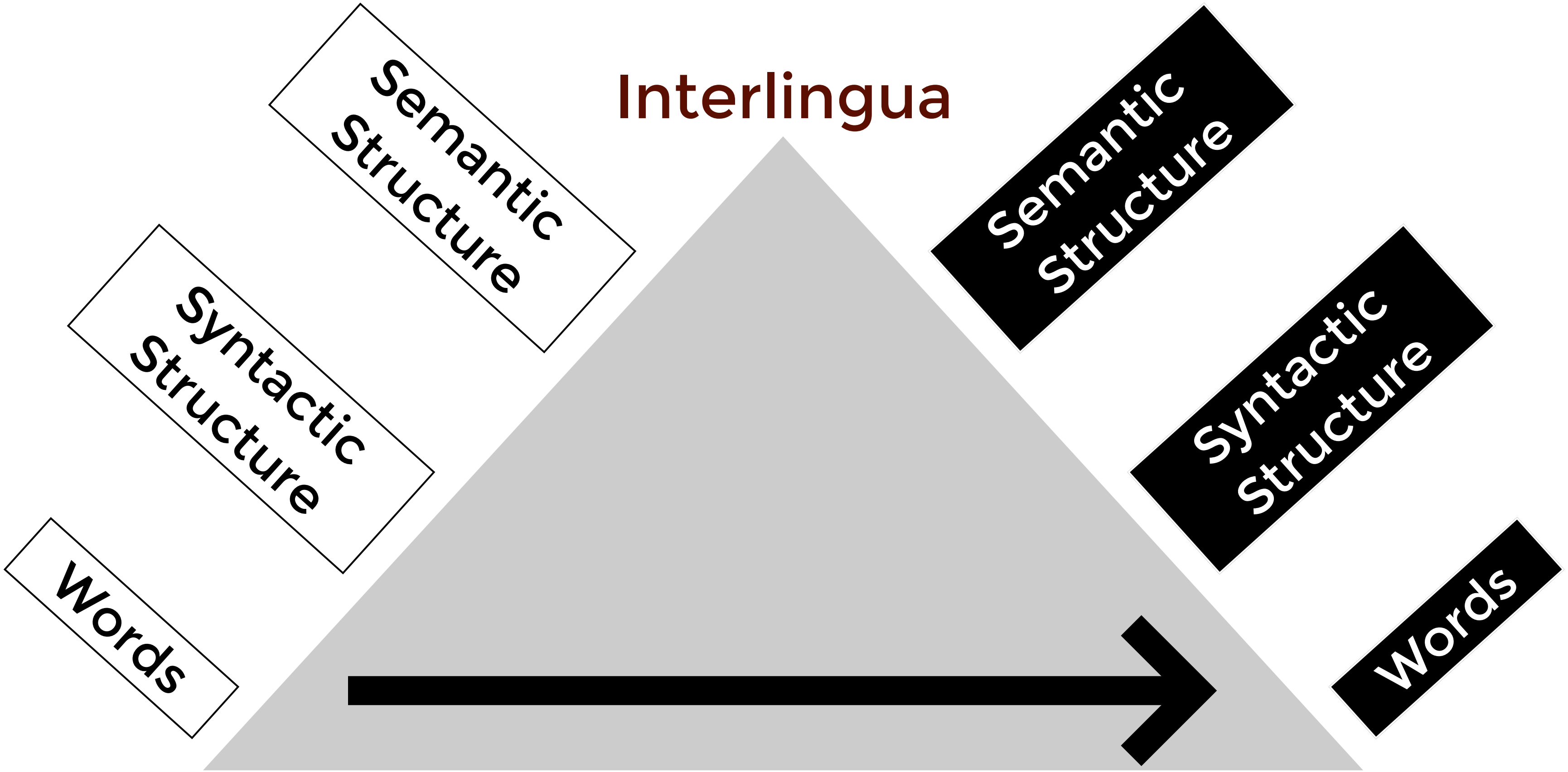
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Tasks in NLP



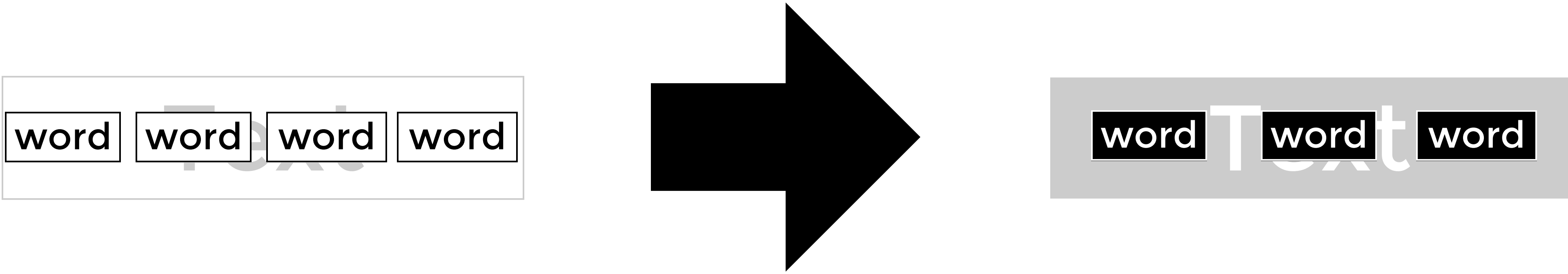
Supervised

Machine Translation



Text-to-text Problem

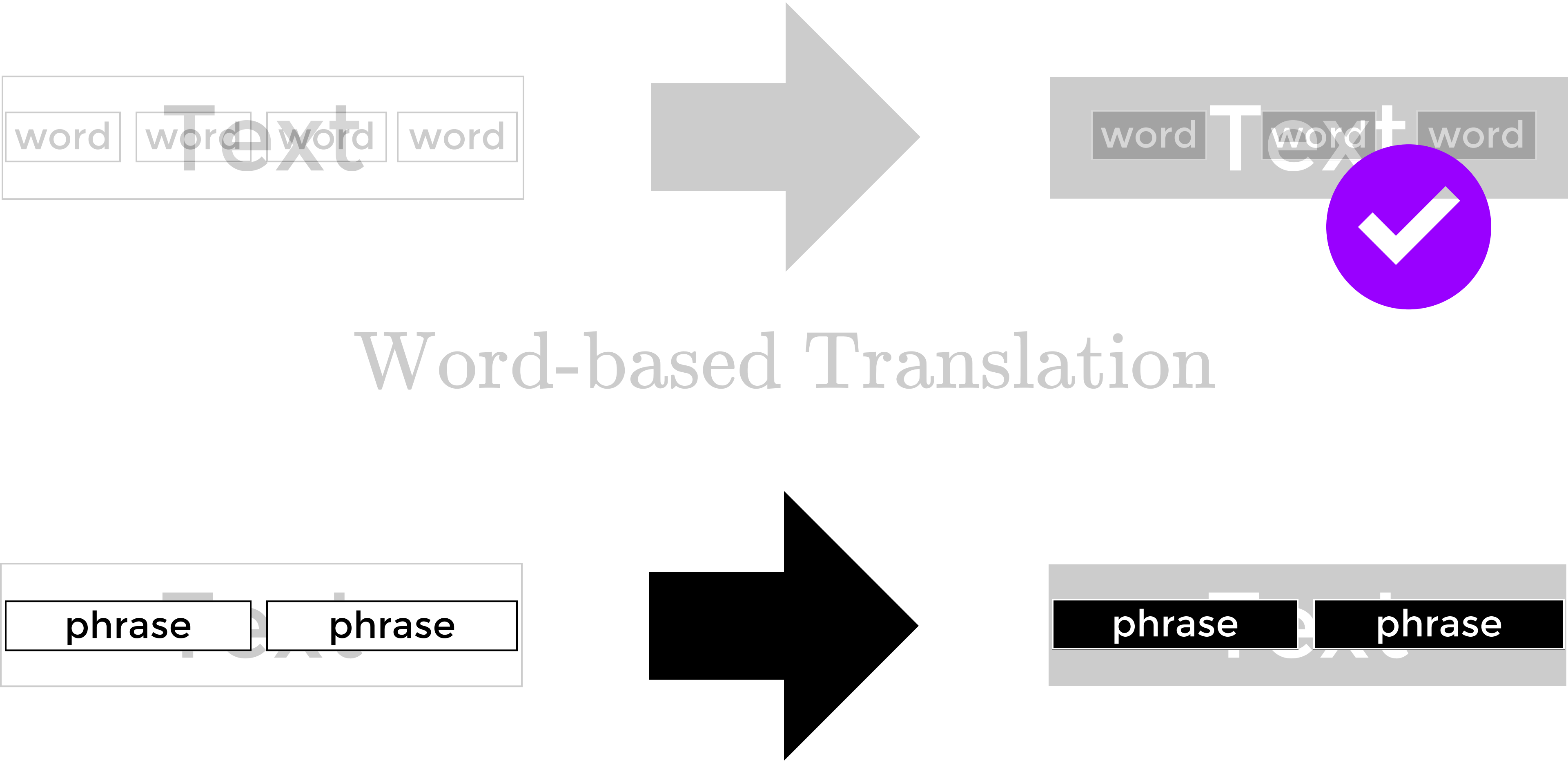
Machine Translation



Word-based Translation

Phrase-based Translation

Machine Translation



Phrase-based Translation

Phrase-based Translation

Optional

The Alignment Template Approach to Statistical Machine Translation

Franz Josef Och*
Google

Hermann Ney†
RWTH Aachen

A phrase-based statistical machine translation approach — the alignment template approach — is described. This translation approach allows for general many-to-many relations between wor[ds]. Thereby, the context of words is taken into account in the translation model, and local change[s] in word order from source to target language can be learned explicitly. The model is described using a log-linear modeling approach, which is a generalization of the often used source–channel approach. Thereby, the model is easier to extend than classical statistical machine translation systems. We describe in detail the process for learning phrasal translations, the feature functions used, and the search algorithm. The evaluation of this approach is performed on three different tasks. For the German–English speech VERBMOBIL *task, we analyze the effect of various system components. On the French–English Canadian* HANSARDS *task, the alignment template system obtains significantly better results than a single-word-based translation model. In the Chinese–English 2002 National Institute of Standards and Technology (NIST) machine translation evaluation it yields statistically significantly better NIST scores than all competing research and commercial translation systems.*

1. Introduction

Machine translation (MT) is a hard problem, because natural languages are highly complex, many words have various meanings and different possible translations, sentences might have various readings, and the relationships between linguistic entities are often vague. In addition, it is sometimes necessary to take world knowledge into account. The number of relevant dependencies is much too large and those dependencies are too complex to take them all into account in a machine translation system. Given these boundary conditions, a machine translation system has to make decisions (produce translations) given incomplete knowledge. In such a case, a principled approach to solving that problem is to use the concepts of statistical decision theory to try to make optimal decisions given incomplete knowledge. This is the goal of statistical machine translation.

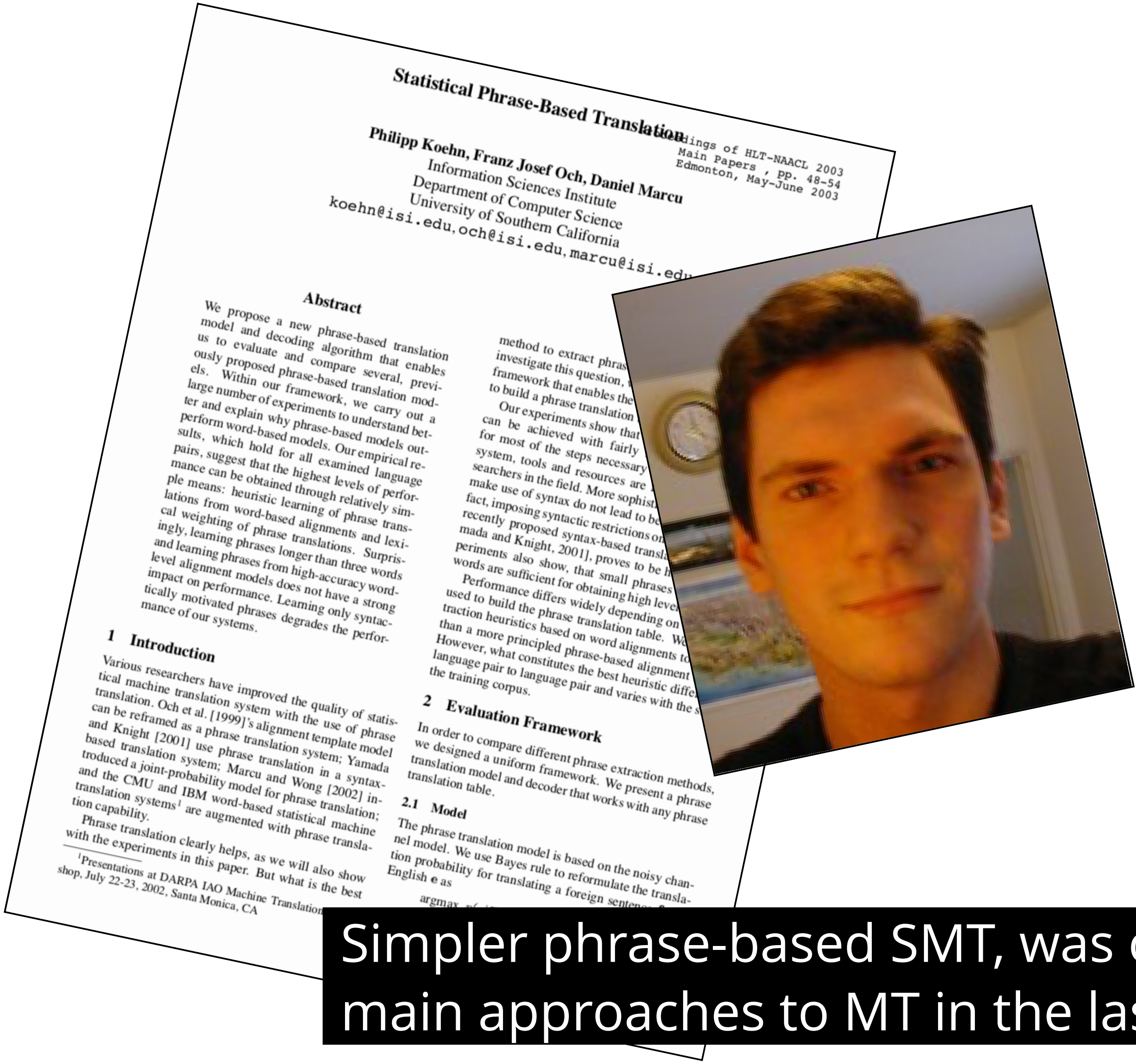
The use of statistical techniques in machine translation has led to dramatic improvements in the quality of research systems in recent years. For example, the statistical approaches of the VERBMOBIL evaluations (Wahlster 2000) or the U.S. National

* 1600 Amphitheatre Parkway, Mountain View, CA 94043. E-mail: och@google.com.
† Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen–University of Technolo
Ahornstr. 55, 52056 Aachen, Germany. E-mail: ney@cs.rwth-aachen.de.

Submission received: 19 November 2002; Revised submission recei
Publication: 1 June 2004

Probably the first successful approach to phrase-based MT, but is complicated!

Phrase-based Translation



Statistical Phrase-Based Translation

Proceedings of HLT-NAACL 2003
Main Papers , pp. 48-54
Edmonton, May-June 2003

Philipp Koehn, Franz Josef Och, Daniel Marcu
Information Sciences Institute
Department of Computer Science
University of Southern California
koehn@isi.edu, och@isi.edu, marcu@isi.edu

Abstract

We propose a new phrase-based translation model and decoding algorithm that enables us to evaluate and compare several, previously proposed phrase-based translation models. Within our framework, we carry out a large number of experiments to understand better and explain why phrase-based models outperform word-based models. Our empirical results, which hold for all examined language pairs, suggest that the highest levels of performance can be obtained through relatively simple means: heuristic learning of phrase translations from word-based alignments and lexical weighting of phrase translations. Surprisingly, learning phrases longer than three words and learning phrases from high-accuracy word-level alignment models does not have a strong impact on performance. Learning only syntactically motivated phrases degrades the performance of our systems.

1 Introduction

Various researchers have improved the quality of statistical machine translation system with the use of phrase translation. Och et al. [1999]'s alignment template model can be reframed as a phrase translation system; Yamada and Knight [2001] use phrase translation in a syntax-based translation system; Marcu and Wong [2002] introduced a joint-probability model for phrase translation; and the CMU and IBM word-based statistical machine translation systems[1] are augmented with phrase translation capability.

Phrase translation clearly helps, as we will also show with the experiments in this paper. But what is the best

method to extract phras
investigate this question,
framework that enables the
to build a phrase translation

Our experiments show that
can be achieved with fairly
for most of the steps necessary
system, tools and resources are
searchers in the field. More sophist
make use of syntax do not lead to be
fact, imposing syntactic restrictions on
recently proposed syntax-based transla
mada and Knight, 2001], proves to be h
periments also show, that small phrases
words are sufficient for obtaining high leve

Performance differs widely depending on
used to build the phrase translation table. We
traction heuristics based on word alignments to
than a more principled phrase-based alignment
However, what constitutes the best heuristic diffe
language pair to language pair and varies with the s
the training corpus.

2 Evaluation Framework

In order to compare different phrase extraction methods, we designed a uniform framework. We present a phrase translation model and decoder that works with any phrase translation table.

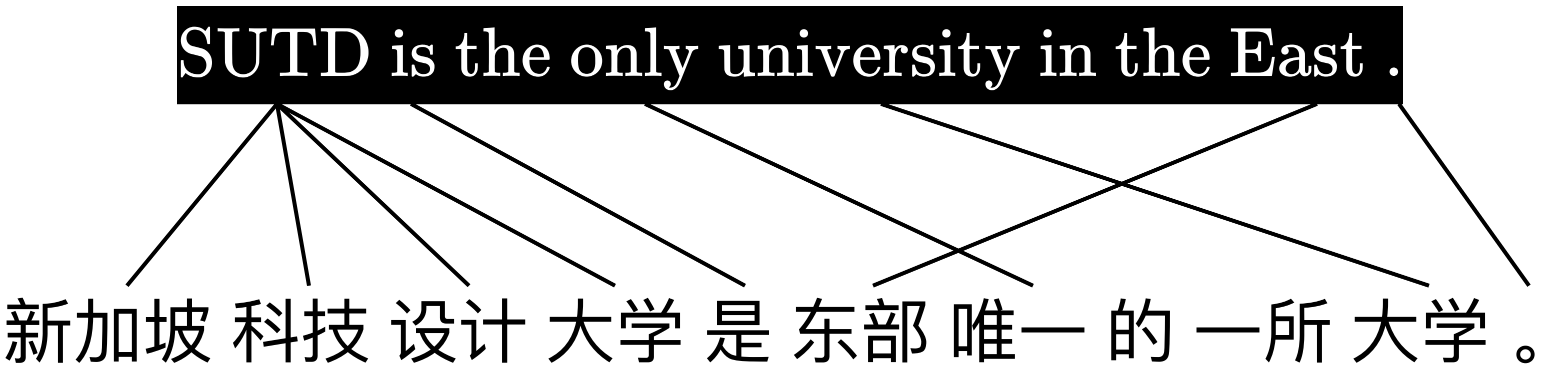
2.1 Model

The phrase translation model is based on the noisy channel model. We use Bayes rule to reformulate the translation probability for translating a foreign sentence f into English **e** as

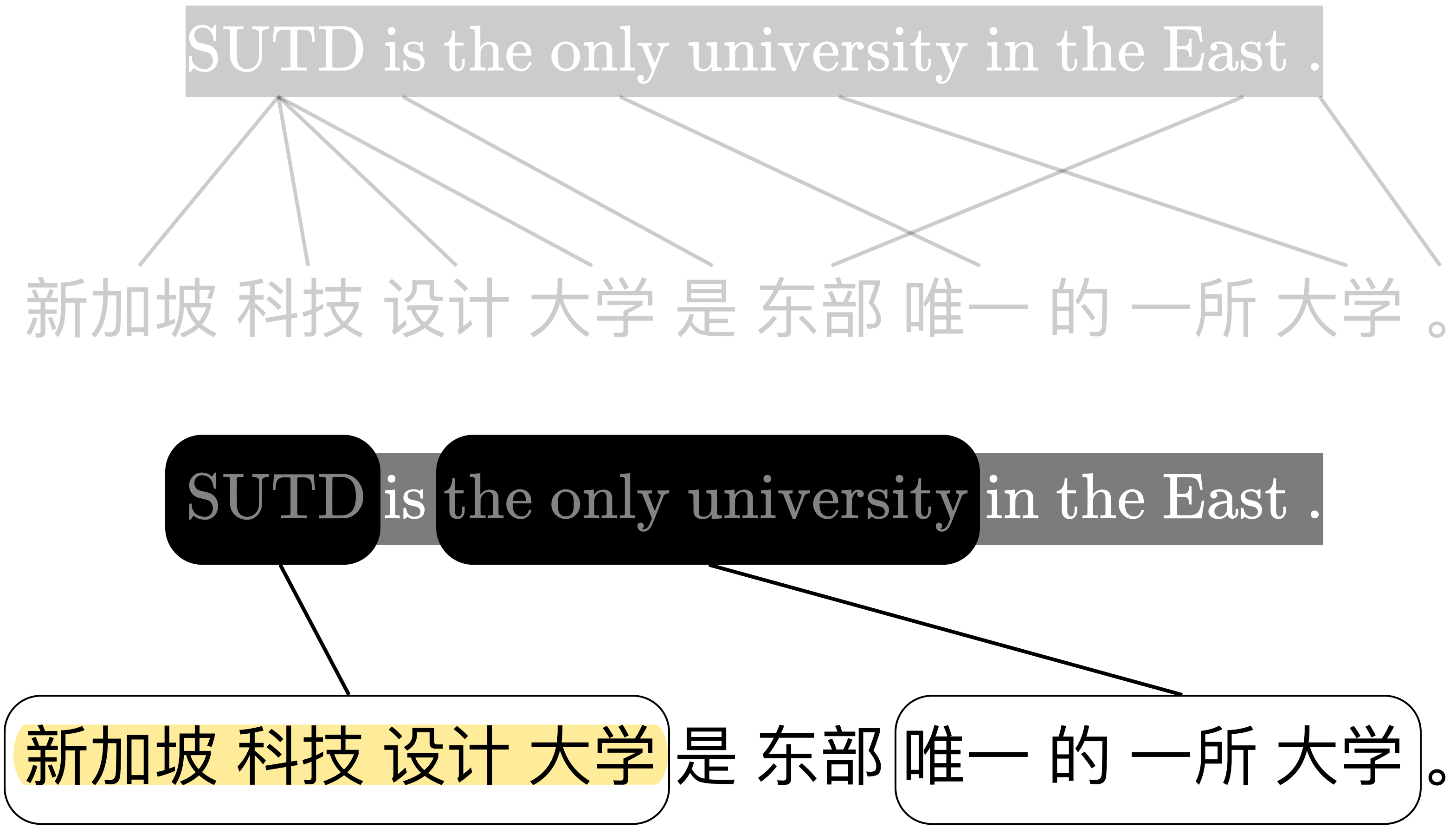
[1] Presentations at DARPA IAO Machine Translation Workshop, July 22-23, 2002, Santa Monica, CA

Simpler phrase-based SMT, was one of the main approaches to MT in the last decade.

Word Alignment



Phrase Alignment

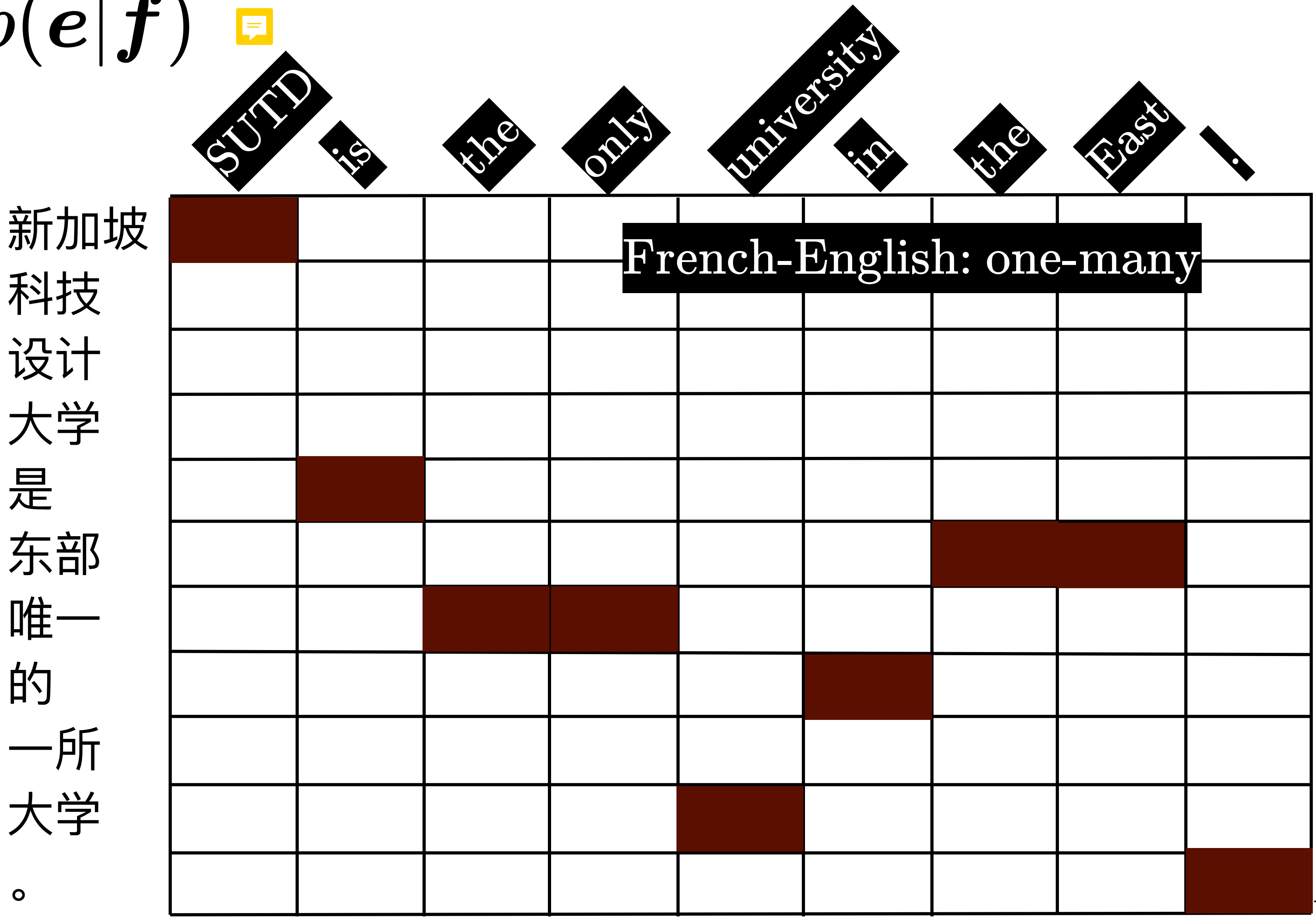


Phrase Lexicon

新加坡 科技 设计 大学	$\Rightarrow$	SUTD
唯一 的 一所 大学	$\Rightarrow$	the only university

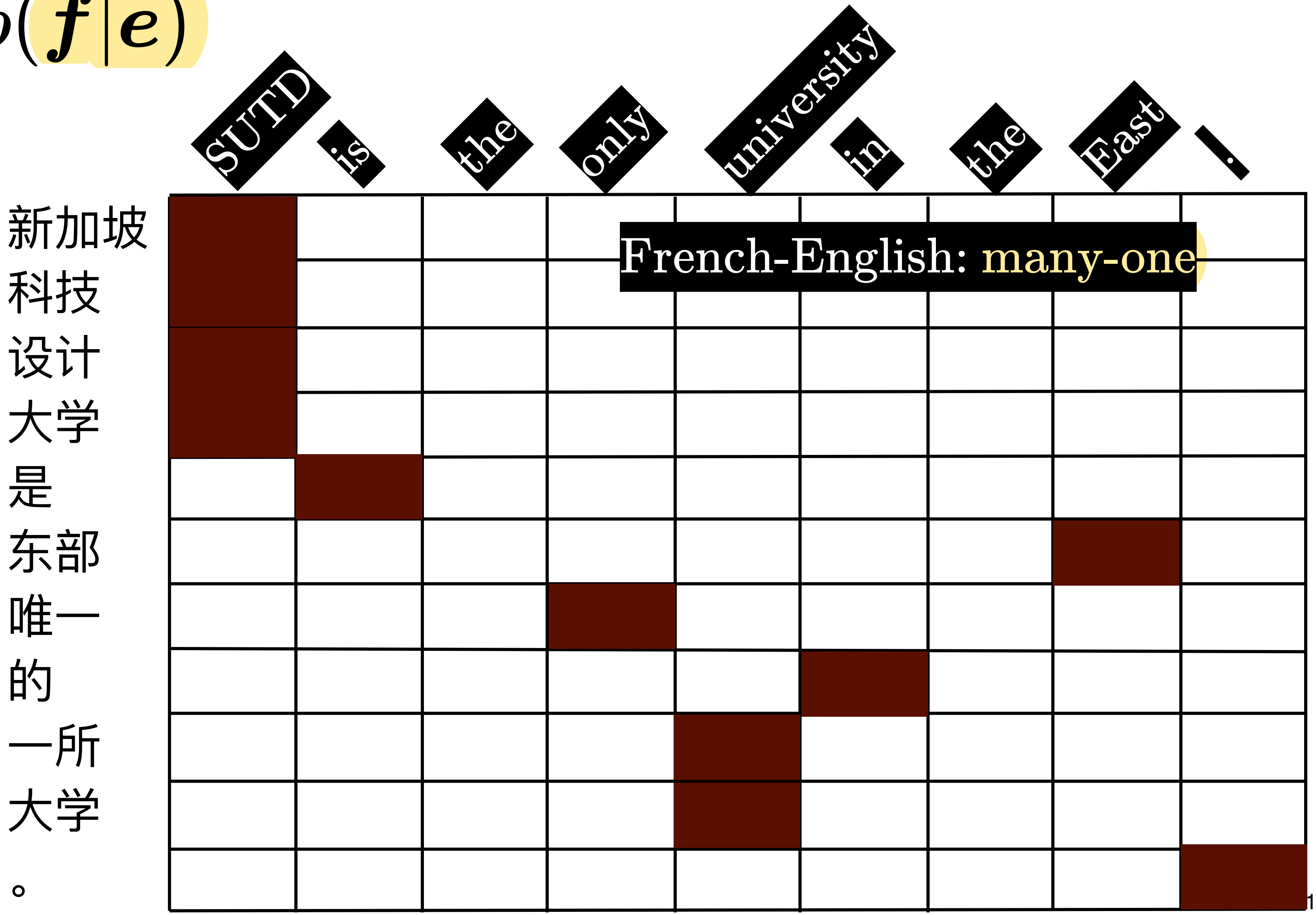
Word Alignment

$\mathbf{A} : p(\boldsymbol{e}|\boldsymbol{f})$



Word Alignment

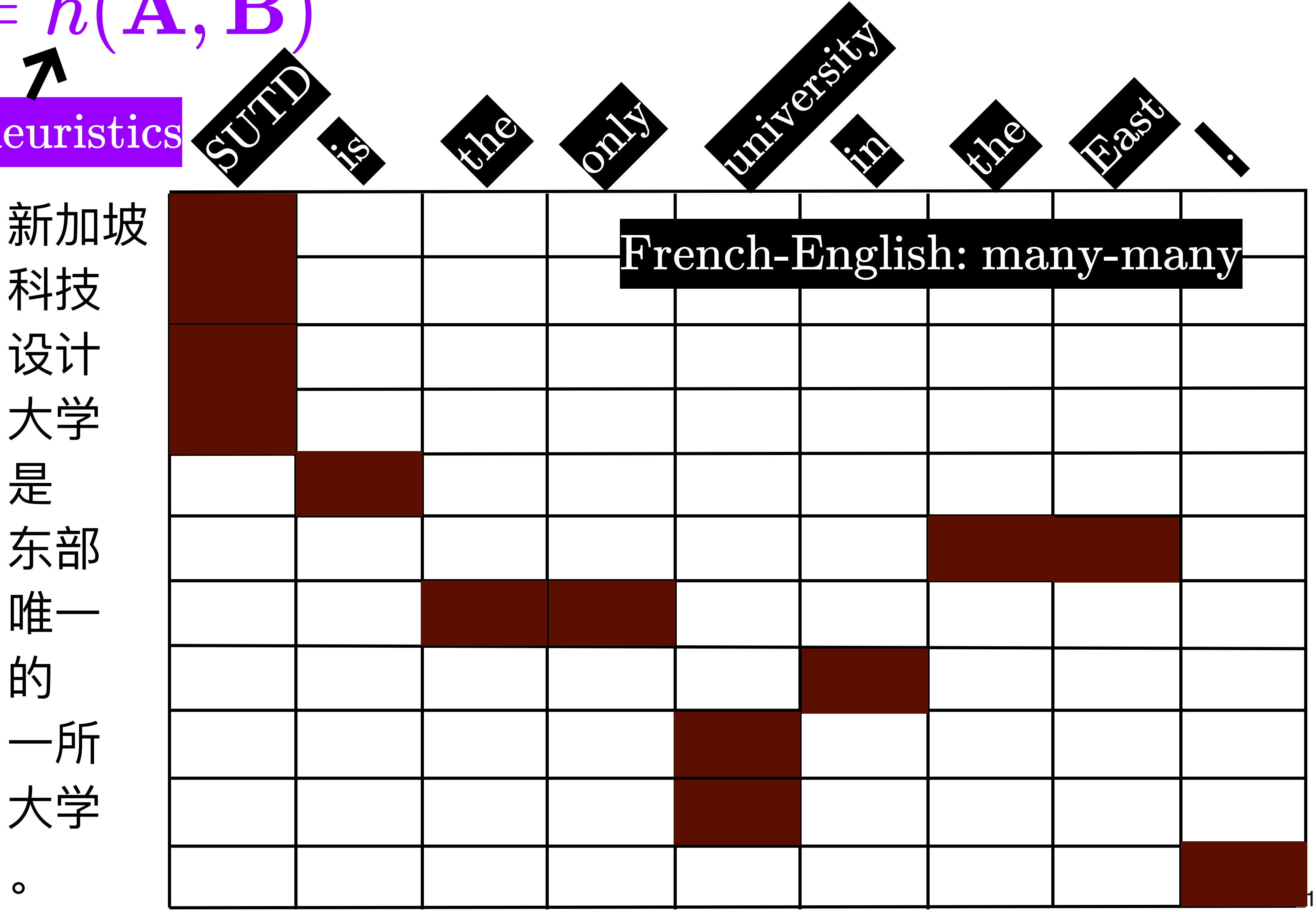
$\mathbf{B}: p(\boldsymbol{f}|\boldsymbol{e})$



Phrase Alignment

$$\mathbf{M} = h(\mathbf{A}, \mathbf{B})$$

some heuristics



Phrase Alignment

$$\mathbf{M} = h(\mathbf{A}, \mathbf{B})$$

Some sample heuristics

Start with the intersection of **A** and **B**

Incrementally add points from union of **A** and **B**

First only add points to words which are not aligned

Give priority to points with neighboring points

Phrase Lexicon

We need an algorithm to extract the phrase pairs

A phrase pair $(\bar{f}, \bar{e})$ is **consistent** if:

At least one word in $\bar{f}$ aligns with a word in $\bar{e}$

No words in $\bar{e}$ align to words outside $\bar{f}$

No words in $\bar{f}$ align to words outside $\bar{e}$

Extract all consistent phrase pairs from the training set

Phrase-based Translation

LM : Language Model

How well the translated sentence reads

TM : Phrase Translation Model

How faithful the translation is to the original

DM : Distortion Model

How much efforts on "moving the eyes" in translation is required

Phrase Translation Model

We need to score each extracted phrase pair

A phrase pair $(\bar{f}, \bar{e})$ can be scored as:

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

This is the "phrase translation probability", which says something about the quality of this translation pair

Distortion Model

A distortion parameter that says the significance of the amount of distortion (usually negative).

Position of the first word in the French phrase that corresponds to the current translated English phrase

$$\eta \times |\mathrm{pl}(p_k) + 1 - \mathrm{pf}(p_{k+1})|$$

Position of the last word in the French phrase that corresponds to the previous translated English phrase

$$\boldsymbol{e} = \underbrace{p_1 p_2 \cdots p_{L-1} p_L}_{L \text{ phrases}}$$

This quantity measures how much efforts on "moving the eyes" is needed when the translator is doing the translation.

Distortion Model

Typically, there is a limit to the maximal distortion we can tolerate.

$$|\mathrm{pl}(p_k) + 1 - \mathrm{pf}(p_{k+1})| \leq d$$

Large distortion can lead to poor translation quality in practice

Phrase-based Translation

$p_1 = (1, 4, \text{SUTD})$

SUTD

$\boxed{\text{新加坡 科技 设计 大学}}$ 是 东部 唯一 的 一所 大学 。

$$\underbrace{\log q(\text{SUTD}|\langle\texttt{START}\rangle, \langle\texttt{START}\rangle)}_{\text{Language Model}}$$

$$+ \quad \underbrace{\log t(\text{新加坡 科技 设计 大学}|\text{SUTD})}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 0}_{\text{Distortion model}}$$

Phrase-based Translation

$p_2 = (5, 5, \text{is})$

SUTD is

新加坡 科技 设计 大学 $\boxed{\text{是}}$ 东部 唯一 的 一所 大学 。

$$\underbrace{\log q(\text{is} | \langle \texttt{START} \rangle, \text{SUTD})}_{\text{Language model}}$$

$$+ \quad \underbrace{\log t(\text{是} | \text{is})}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 0}_{\text{Distortion model}}$$

Phrase-based Translation

$p_3 = (7, 8, \text{the only})$

SUTD is the only

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

$$\underbrace{\log q(\text{the}|\text{SUTD}, \text{is}) + \log q(\text{only}|\text{is}, \text{the})}_{\text{Language model}}$$

$$+ \quad \underbrace{\log t(\text{唯一 的}|\text{the only})}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 1}_{\text{Distortion model}}$$

Phrase-based Translation

$p_4 = (9, 10, \text{university})$

SUTD is the only university

新加坡 科技 设计 大学 是 东部 唯一 的 $\boxed{\text{一所 大学}}$ 。

$$\underbrace{\log q(\text{university}|\text{the, only})}_{\text{Language model}}$$

$$+ \quad \underbrace{\log t(\text{一所 大学}|\text{university})}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 0}_{\text{Distortion model}}$$

Phrase-based Translation

$p_5 = (6, 6, \text{in the East})$

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 。

$$\underbrace{\log q(\text{in}|\text{only, university}) + \log q(\text{the}|\text{university, in}) + \log q(\text{East}|\text{in, the})}_{\text{Language model}}$$

$$+ \quad \underbrace{\log t(\text{东部}|\text{in the East})}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 5}_{\text{Distortion model}}$$

Phrase-based Translation

$p_6 = (11, 11, .)$

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 $\boxed{。}$

$$\underbrace{\log q(.|\text{the}, \text{East})}_{\text{Language model}}$$

$$+ \quad \underbrace{\log t(。|.)}_{\text{Phrase translation model}}$$

$$+ \quad \underbrace{\eta \times 4}_{\text{Distortion model}}$$

Phrase-based Translation

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 。

$$\text{score}(\boldsymbol{e}) = \underbrace{\log q(\boldsymbol{e})}_{\text{Language model}}$$

$$+ \underbrace{\sum_{k=1}^{L} \log t(p_k)}_{\text{Phrase translation model}}$$

$$+ \underbrace{\sum_{k=1}^{L-1} \eta \times |\text{pl}(p_k) + 1 - \text{pf}(p_{k+1})|}_{\text{Distortion model}}$$

Decoding

We know how to score a translation derivation, but how do we search for the most optimal derivation?

The position of the last French word in the previous French phrase translated

The score of the partial derivation so far

A state

$$s = (e_1, e_2, b, r, \alpha)$$

The last two English words in the previous translated English phrase

A bit string indicating which words in French are (not yet) translated.

States

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

$$(\langle \texttt{START} \rangle, \langle \texttt{START} \rangle, 0000000000, 0, 0)$$

Initial State

States

$p_1 = (1, 4, \mathrm{SUTD})$

SUTD

$\boxed{\text{新加坡 科技 设计 大学}}$ 是 东部 唯一 的 一所 大学 。

$$(\langle \texttt{START} \rangle, \mathrm{SUTD}, 1111000000, 4, 3.7)$$

States

$p_2 = (5, 5, \text{is})$

SUTD is

新加坡 科技 设计 大学 $\boxed{\text{是}}$ 东部 唯一 的 一所 大学 。

$$(\text{SUTD}, \text{is}, 1111000000, 5, 8.9)$$

States

$p_3 = (7, 8, \text{the only})$

SUTD is the only

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

$$(\text{the}, \text{only}, 11111011000, 8, -0.9)$$

States

$$p_4 = (9, 10, \text{university})$$

SUTD is the only university

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

$$(\text{only}, \text{university}, 11111011110, 10, 2.2)$$

States

$p_5 = (6, 6, \text{in the East})$

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 。

$$(\text{the}, \text{East}, 11111111110, 6, 7.1)$$

States

$p_6 = (11, 11, .)$

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 。

$$(\text{East}, ., 11111111111, 11, 5.0)$$

One Final State

States

$p_6 = (11, 11, .)$

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 。

It is similar to a transition-based parser!

$(\text{East}, ., 1111111111, 11, 5.0)$

One Final State

State Transition

$p_2 = (5, 5, \text{is})$

SUTD is

新加坡 科技 设计 大学 是 东部 唯一 的 一所 大学 。

Each p_k is essentially an action!

$$(\langle \texttt{START} \rangle, \text{SUTD}, 1111100000, 5, 8.9)$$

$$(\ldots, 1111100010, \ldots) \qquad \ldots \qquad (\ldots, 1111100111, \ldots)$$

State Transition

$p_2 = (5, 5, \text{is})$

SUTD is

新加坡 科技 设计 大学 $\boxed{\text{是}}$ 东部 唯一 的 一所 大学 。

Allowable actions

1. The action p_k must be compatible with b.
2. The distortion limit d must be respected.

$(\ldots, 11111000010, \ldots) \quad \ldots \quad (\ldots, 11111000111, \ldots)$

State Transition

$p_2 = (5, 5, \text{is})$

SUTD is

新加坡 科技 设计 大学 $\boxed{是}$ 东部 唯一 的 一所 大学 。

The greedy procedure may not yield the best translation.
Beam search is typically used in practice,
which can also be used for "top-k" decoding.

Beam search: instead of committing to a single next action,
we explore a few options at each point in the search process.

Weighted Score

$$
\begin{aligned}
\text{score}(\boldsymbol{e}) = \quad & \lambda_{\text{LM}} \times \underbrace{\log q(\boldsymbol{e})}_{\text{Language model}} \\
+ \quad & \lambda_{\text{TM}} \times \underbrace{\sum_{k=1}^{L} \log t(p_k)}_{\text{Phrase translation model}} \\
+ \quad & \lambda_{\text{DM}} \times \underbrace{\sum_{k=1}^{L-1} \eta \times |\text{pl}(p_k) + 1 - \text{pf}(p_{k+1})|}_{\text{Distortion model}}
\end{aligned}
$$

Tunable hyper-parameters

Question

How to tune the hyper-parameters?

We shall tune the hyper-parameters to optimize some evaluation metric!

BLEU

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{papineni,roukos,toddward,weijing}@us.ibm.com

Abstract

Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that can not be reused. We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation, and that has little marginal cost per run. We present this method as an automated understudy to skilled human judges which substitutes for them when there is need for quick or frequent evaluations.[1]

1 Introduction

1.1 Rationale

Human evaluations of machine translation (MT) weigh many aspects of translation, including *adequacy*, *fidelity*, and *fluency* of the translation (Hovy, 1999; White and O'Connell, 1994). A comprehensive catalog of MT evaluation techniques and their rich literature is given by Reeder (2001). For the most part, these various human evaluation approaches are quite expensive (Hovy, 1999). Moreover, they can take *weeks* or *months* to finish. This is a big problem because developers of machine translation systems need to monitor the effect of *daily* changes to their systems in order to weed out bad ideas from good ideas. We believe that MT progress stems from evaluation and that there is a logjam of fruitful research ideas waiting to be released from

[1]So we call our method the bilingual evaluation understudy, BLEU.

the evaluation bottleneck. D
fit from an inexpensive automa
quick, language-independent, a
with human evaluation. We prop
tion method in this paper.

1.2 Viewpoint

How does one measure translation p
The closer a machine translation is to a p
human translation, the better it is. This i
tral idea behind our proposal. To judge the
of a machine translation, one measures its cl
to one or more reference human translations ac
ing to a numerical metric. Thus, our MT evalua
system requires two ingredients:

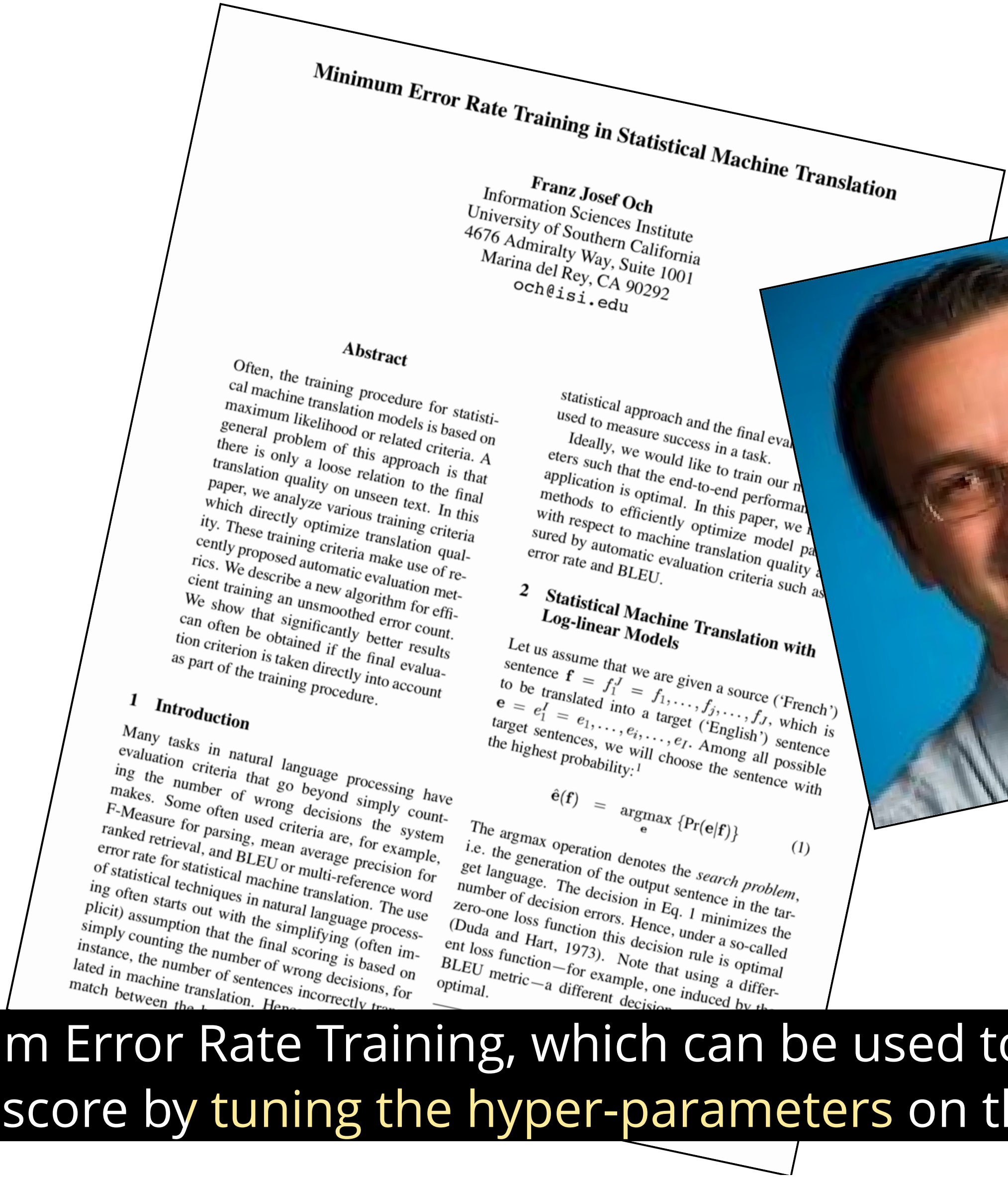
1. a numerical "translation closeness" metric
2. a corpus of good quality human reference translations

We fashion our closeness metric after the highly successful *word error rate* metric used by the speech recognition community, appropriately modified for multiple reference translations and allowing for legitimate differences in word choice and word order. The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric f

In Section 2
detail. In Sectio
BLEU. In Sectio
experiment. In S
metric performan

The most widely adopted evaluation metric for measuring MT quality.

MERT



Minimum Error Rate Training in Statistical Machine Translation

Franz Josef Och
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
och@isi.edu

Abstract

Often, the training procedure for statistical machine translation models is based on maximum likelihood or related criteria. A general problem of this approach is that there is only a loose relation to the final translation quality on unseen text. In this paper, we analyze various training criteria which directly optimize translation quality. These training criteria make use of recently proposed automatic evaluation metrics. We describe a new algorithm for efficient training an unsmoothed error count. We show that significantly better results can often be obtained if the final evaluation criterion is taken directly into account as part of the training procedure.

1 Introduction

Many tasks in natural language processing have evaluation criteria that go beyond simply counting the number of wrong decisions the system makes. Some often used criteria are, for example, F-Measure for parsing, mean average precision for ranked retrieval, and BLEU or multi-reference word error rate for statistical machine translation. The use of statistical techniques in natural language processing often starts out with the simplifying (often implicit) assumption that the final scoring is based on simply counting the number of wrong decisions, for instance, the number of sentences incorrectly translated in machine translation. Hence ... match between the ...

statistical approach and the final eval... used to measure success in a task.

Ideally, we would like to train our m... eters such that the end-to-end performan... application is optimal. In this paper, we ... methods to efficiently optimize model pa... with respect to machine translation quality a... sured by automatic evaluation criteria such as ... error rate and BLEU.

2 Statistical Machine Translation with Log-linear Models

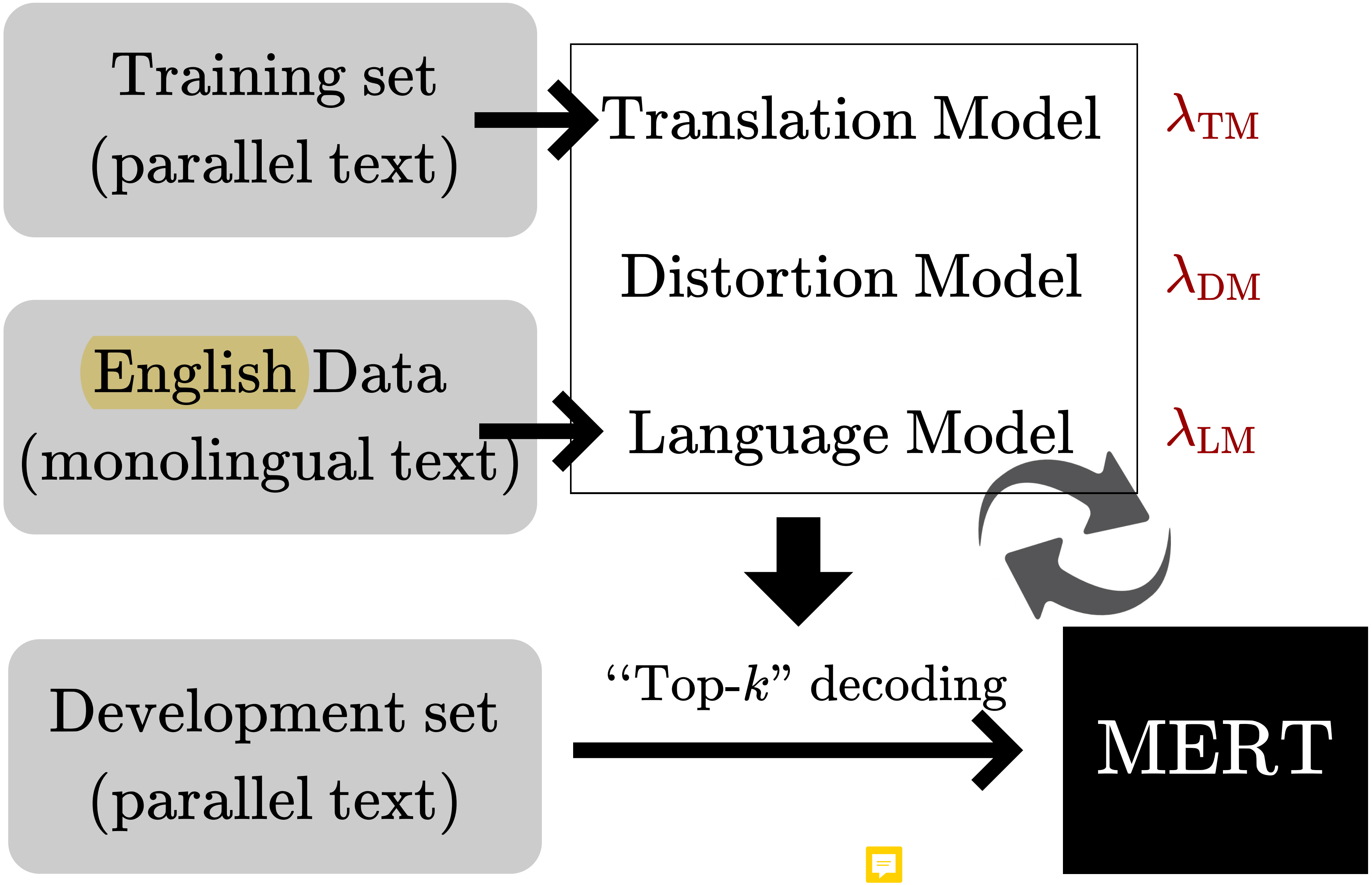
Let us assume that we are given a source ('French') sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$, which is to be translated into a target ('English') sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Among all possible target sentences, we will choose the sentence with the highest probability:[1]

$$\hat{\mathbf{e}}(\mathbf{f}) = \operatorname*{argmax}_{\mathbf{e}} \{\Pr(\mathbf{e}|\mathbf{f})\} \quad (1)$$

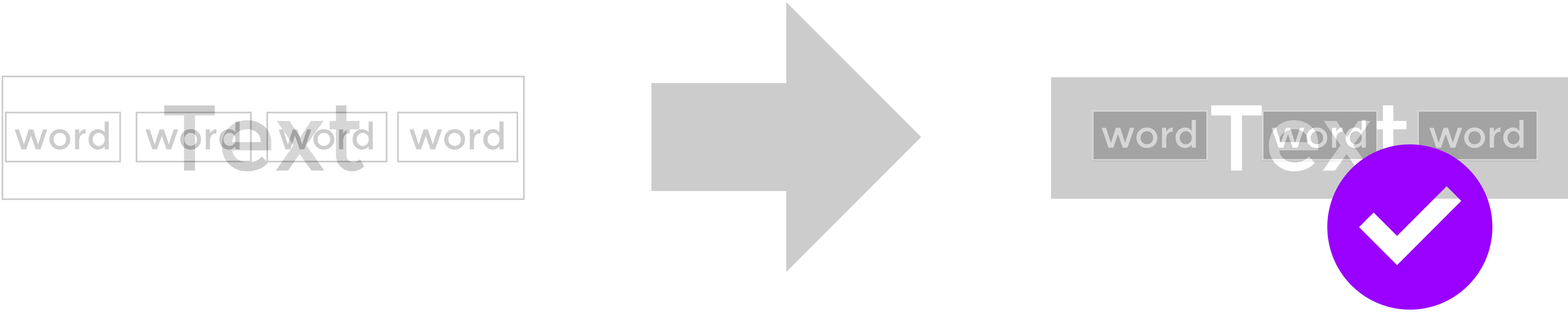
The argmax operation denotes the *search problem*, i.e. the generation of the output sentence in the target language. The decision in Eq. 1 minimizes the number of decision errors. Hence, under a so-called zero-one loss function this decision rule is optimal (Duda and Hart, 1973). Note that using a different loss function—for example, one induced by the BLEU metric—a different decision ... optimal.

Minimum Error Rate Training, which can be used to directly optimize the BLEU score by tuning the hyper-parameters on the development set.

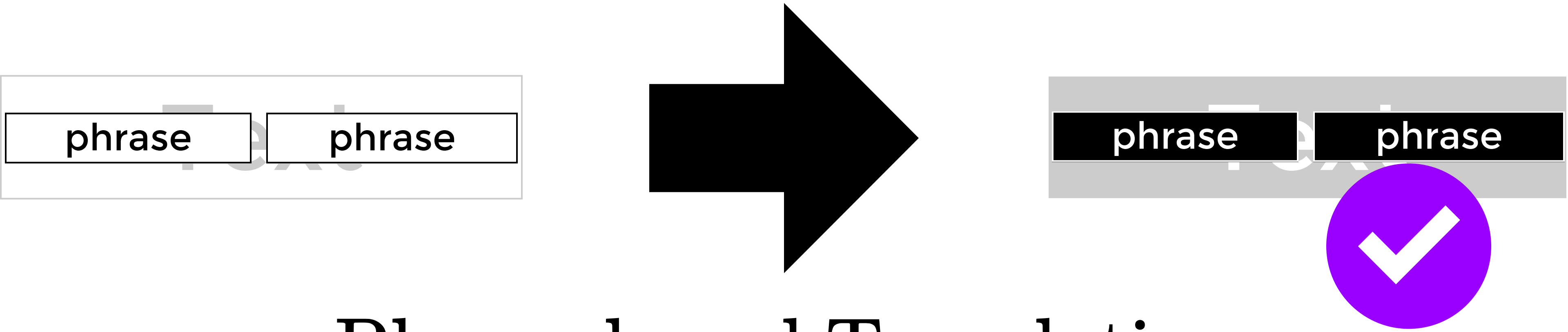
Phrase-based Translation



Machine Translation



Word-based Translation



Phrase-based Translation

Phrase-based Translation

The process involves a transition-based procedure, which was introduced when discussing parsing.

Text

Parser??

Text

Is it possible to involve a parser in the translation process in some way?

Machine Translation