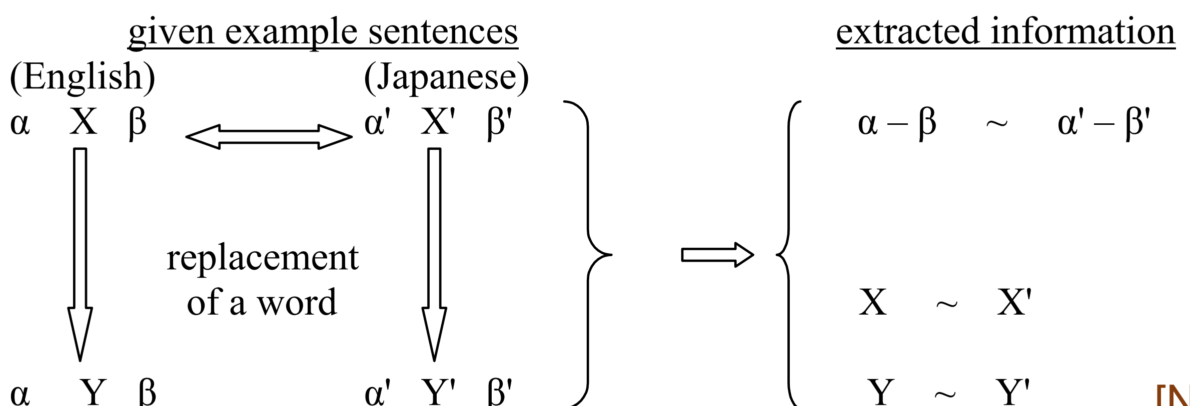
机器翻译的研究历程
--统计机器翻译

黄书剑

基于实例的机器翻译（since 1980s）

- **从语料库中学习翻译实例**
  - 查找接近的翻译实例，并进行逐词替换进行翻译
  - 利用类比思想analogy，避免复杂的结构分析



[Nagao 1984]

  - 例如：

    我 来自 南京 大学《==》I come from Nanjing University

    我 来自 X ~ I come from X
    南京 大学 ~ Nanjing University

机器翻译的发展

机器翻译概念
Weaver, 1949

规则机器翻译
since 1950s

实例机器翻译
Nagao, 1980s

统计机器翻译
Brown et al., 1993
Koehn et al., 2003
Chiang et al., 2005

神经网络机器翻译
Cho et al., 2014
Bahdanau et al., 2015

数据资源越来越丰富
计算能力越来越强

统计机器翻译（since 1990s）

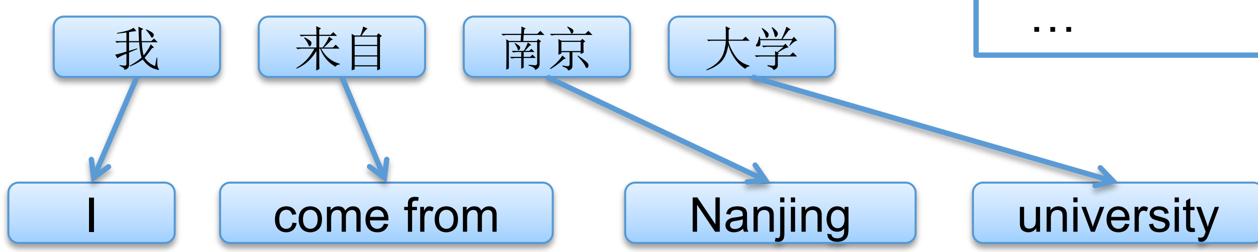
- 从双语平行语料中自动进行翻译规则的学习和应用

我 来自 中国
I come from China

南京 是 六朝 古都 …
Nanjing is the ancient capital of …

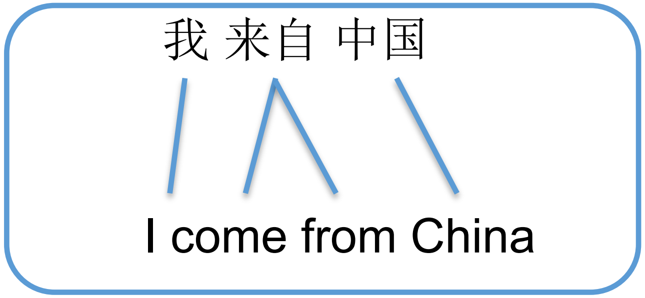
…是一座世界一流大学
…is a world-class university

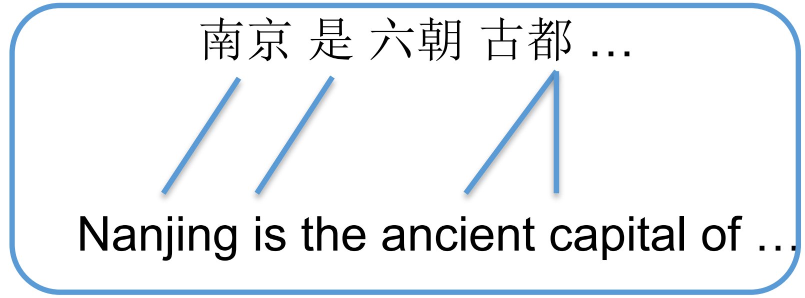
我	I
来自	come from
中国	China
南京	Nanjing
是	is
古都	the ancient capital
一座	a
世界一流	world-class
大学	university
…	…



词对齐

- 自动学习翻译对应关系（词级别的对应）





  - 没有大规模标记数据，采用无监督方法学习
  - 从词语的共现中发掘翻译关系

The Mathematics of Statistical Machine Translation: Parameter Estimation

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

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

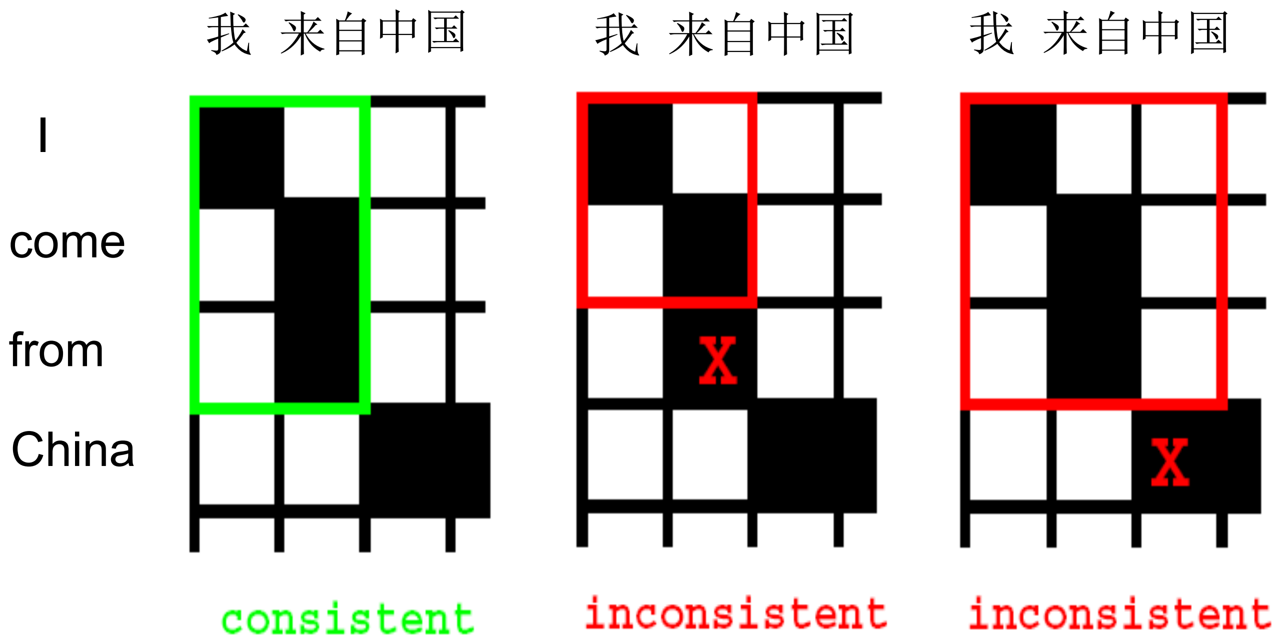
Vincent J. Della Pietra*
IBM T.J. Watson Research Center

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

[Brown et al. 1993]

短语规则

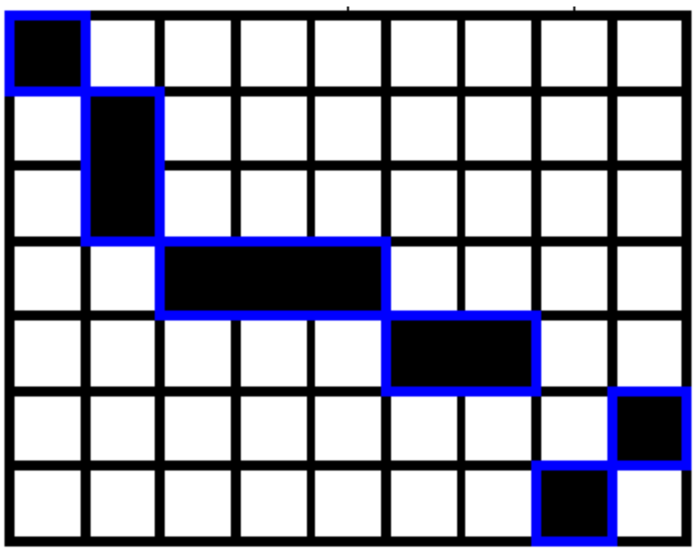
- **根据对应关系抽取翻译规则（短语）**

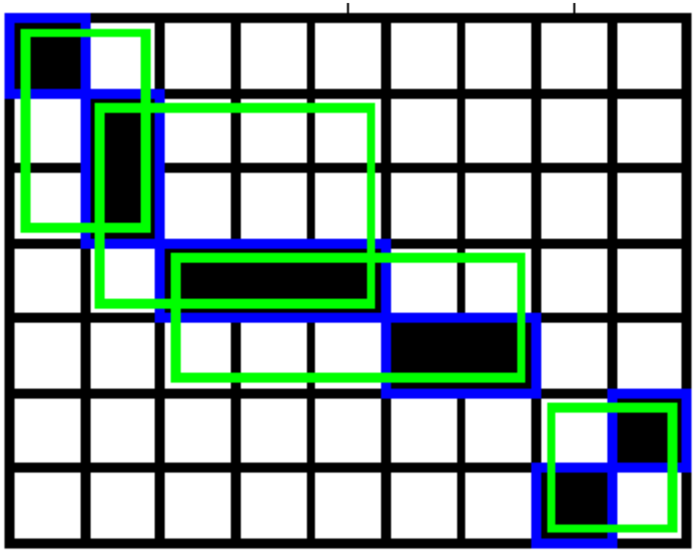


[Koehn et al. 2003]

短语规则的抽取

- **短语关系可能是嵌套的**
  - 六朝、古都、六朝古都
  - 南京、南京大学



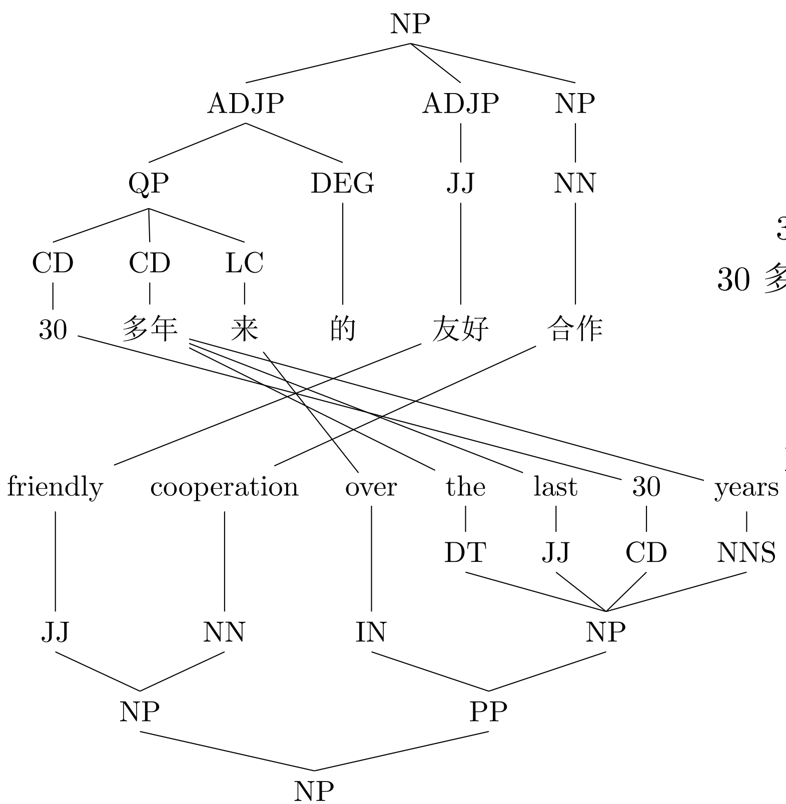


短语规则的评分

- **同一个源语言片段可能存在多个翻译**
  - 如： “来自” 可能被翻译为：come from， came from, is from, are from …
  - 需要对同一个源语言片段的不同翻译进行评价
- **考虑以下一些因素：**
  - 相对频率；
  - 词汇翻译概率；
  - 反向相对频率；
  - 反向词汇翻译概率；

$$\phi(\overline{f}|\overline{e}) = \frac{\text{count}(\overline{f},\overline{e})}{\sum_{\overline{f}} \text{count}(\overline{f},\overline{e})}$$

更复杂的翻译模型



(a) 双语句对、句法树及词对齐

30–>30　　多年–>the, last, years

来–>over　　友好–>friendly

(b) 单词翻译规则示例

30 多年–>the last 30 years　　友好 合作–>friendly cooperation

30 多年 来–>over the last 30 years　　的 友好–>friendly

(c) 短语翻译规则示例

30–>30　　X 的 X–>X2 X1

X 多年–>the last X years　　友好 合作–>friendly cooperation

(d) 层次翻译规则示例

QP(CD 30)(CD 多年)(LC 来)–>the last 30 years

友好 合作–>NP(JJ friendly)(NN cooperation)

QP(CD 30)(CD 多年)(LC 来)–>NP(DT the)(JJ last)(CD 30)(NNS years)

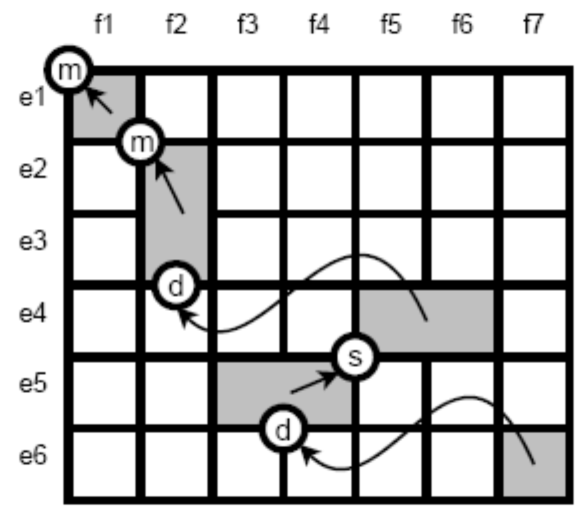
(e) 句法翻译规则示例

[Chiang 2005 ACL best paper, Yamada and Knight 2001, Liu et al. 2006]

翻译顺序的调整

- **顺序翻译Monotone translation**
  – 不允许任何顺序变化
- **基于距离的调序限制Distance-based reordering cost**
  – 根据顺序调整的长度进行惩罚
- **词汇化的调序模型:**
  - Monotone
  - Swap
  - Discontinuous

  – 由双语的单词和短语来决定调序
  - Eg: P(Mono | no, did not)



- **通过语言模型来建模句子的流畅程度**

生成式模型中的参数

- **生成式模型：** The Source-Channel Model [Brown et al. 1993]
  - 参数即概率模型

$$Pr(\mathbf{e}|\mathbf{f}) = Pr(\mathbf{e})Pr(\mathbf{f}|\mathbf{e})/Pr(\mathbf{f})$$

语言模型　　翻译模型

  - 概率模型的估计采用训练数据上的极大似然估计进行

判别式模型中的参数

- 判别式模型：Log-linear Models
  - [Och and Ney, 2002 ACL best paper]
  - 参数包括子模型参数及对数线性模型参数

$$Pr(\mathbf{e}|\mathbf{f}) = p_{\lambda_1^M}(\mathbf{e}|\mathbf{f})$$

$$= \frac{exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}|\mathbf{f})]}{\sum_{\mathbf{e}'} exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}'|\mathbf{f})]}$$

对数线性模型的参数☆

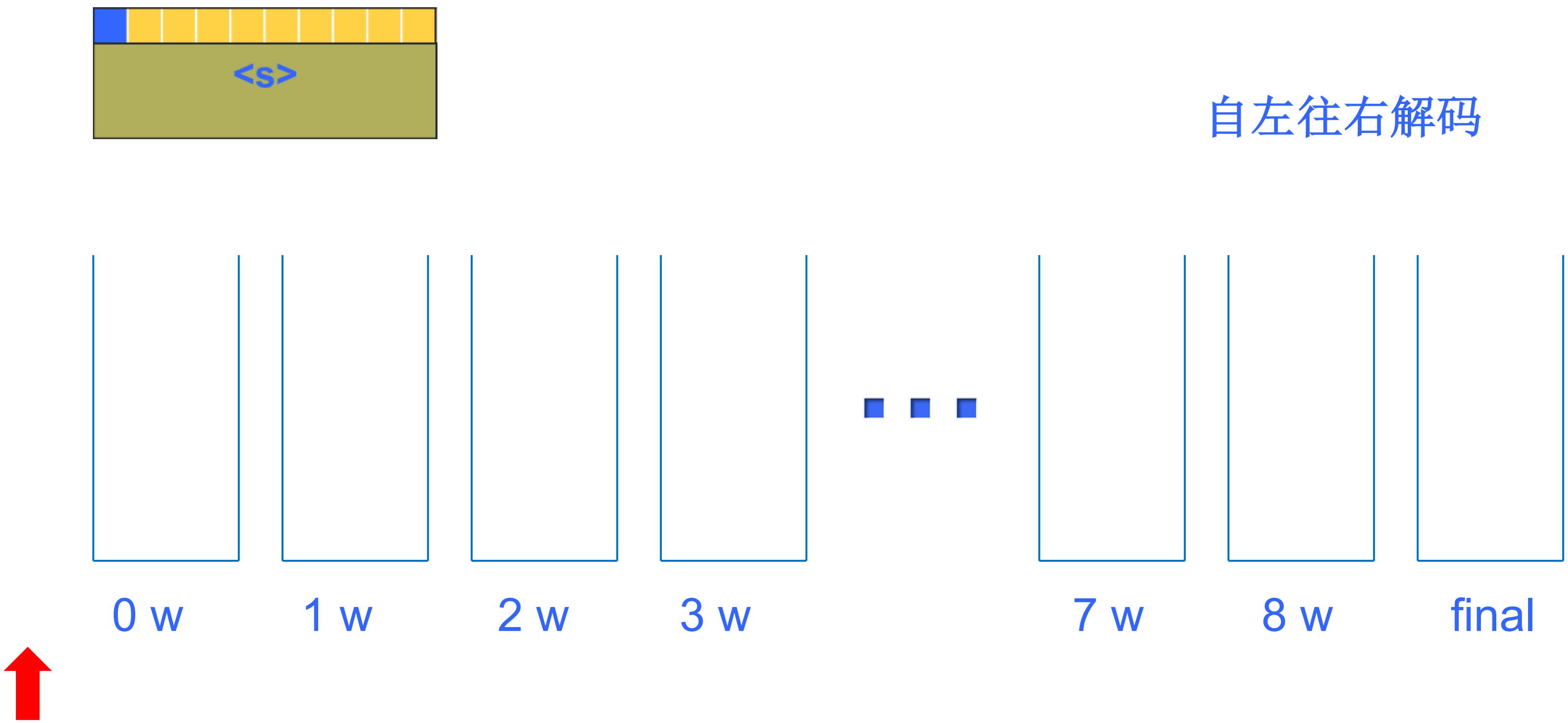
子模型参数：短语翻译模型
词翻译模型
调序模型
语言模型
…

翻译搜索/解码

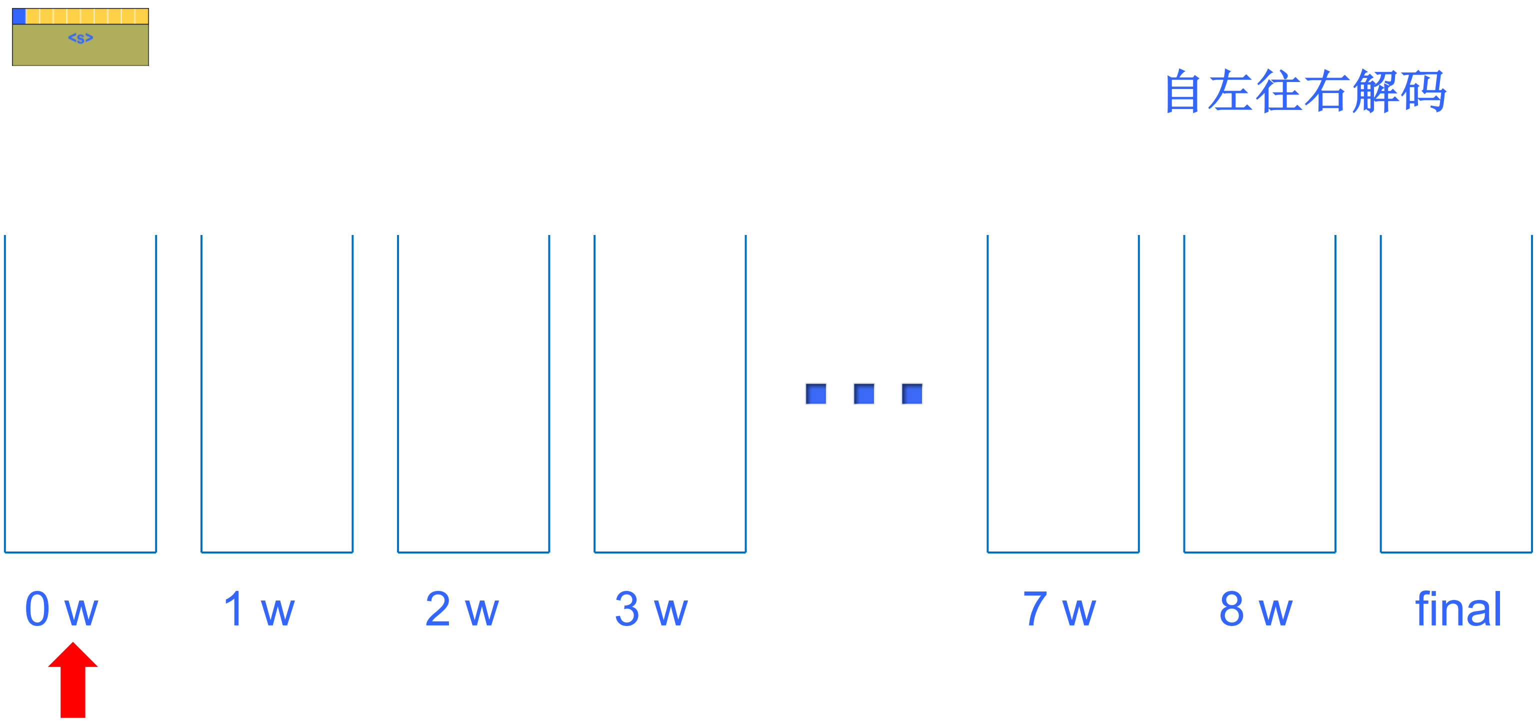
- 根据前述的模型和参数查找最优的翻译结果
  - non-local模型/评分函数
- 从所有可能的翻译结果中找出“最优解”
  - 指数级的搜索空间
  - 如何比较两个翻译候选?
    - 翻译过的词不同、翻译选择不同

翻译搜索/解码

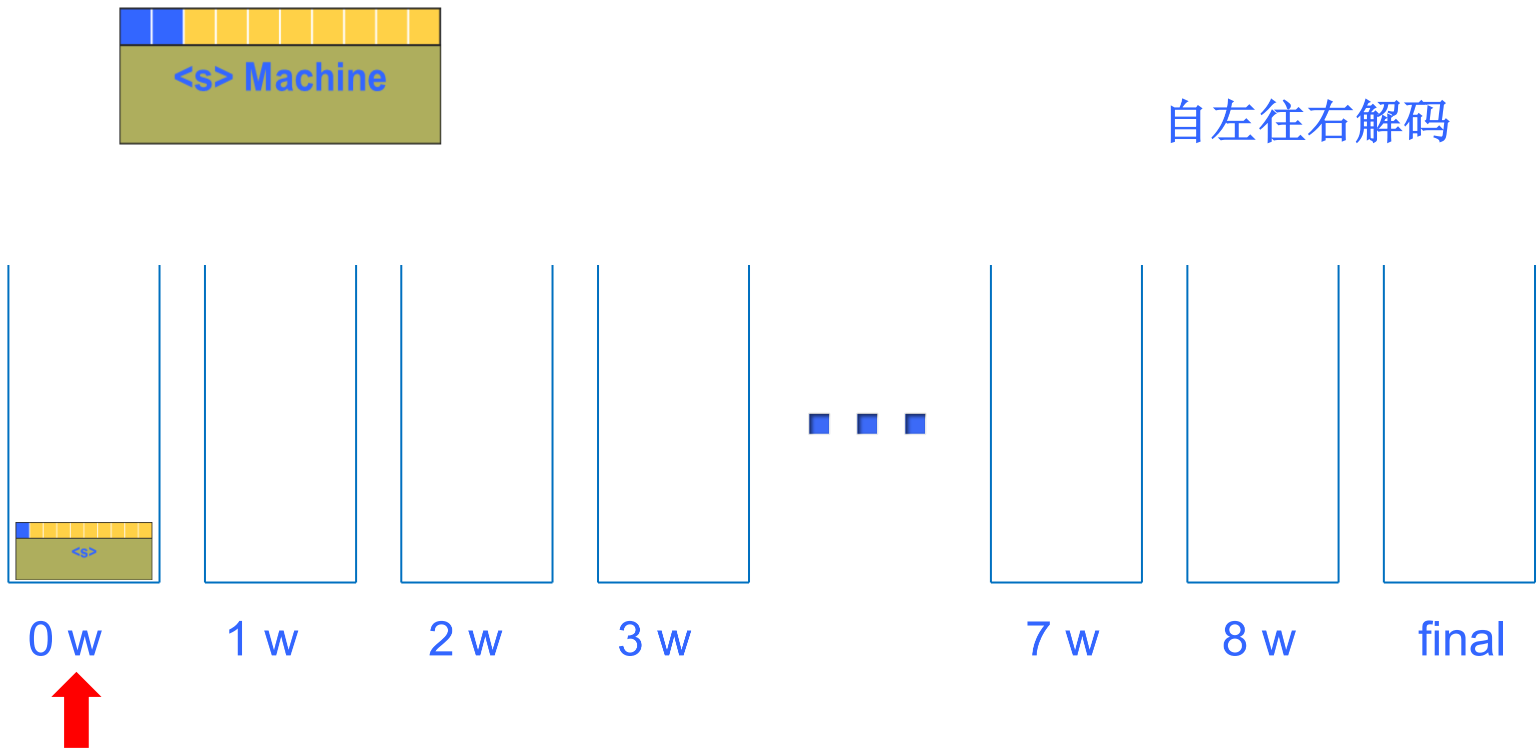
<s> 机 器 翻 译 是 个 崭 新 的 领 域 。</s>



<s>机器 翻译 是 个 崭新 的 领域 。</s>

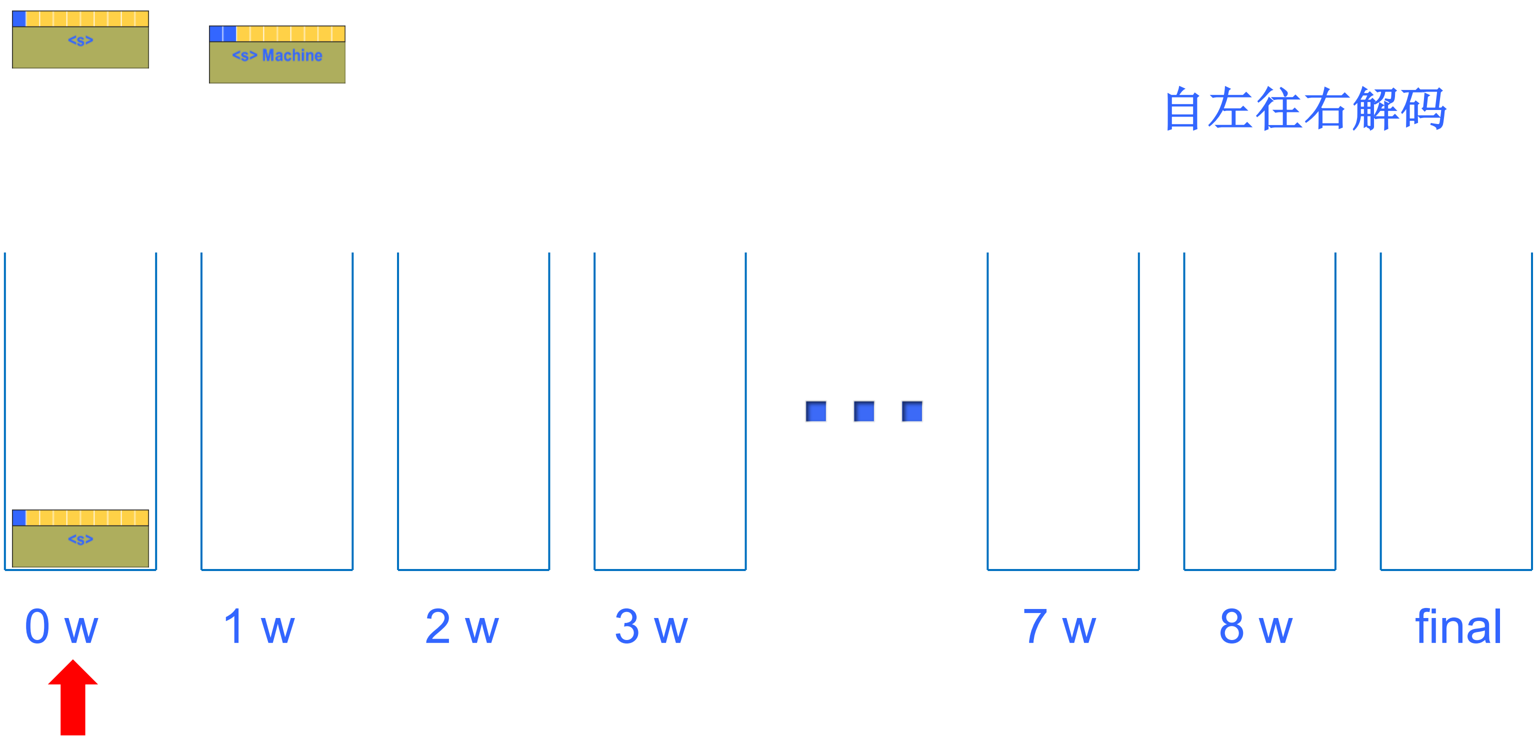


<s> 机器 翻译 是 个 崭新 的 领域 。</s>

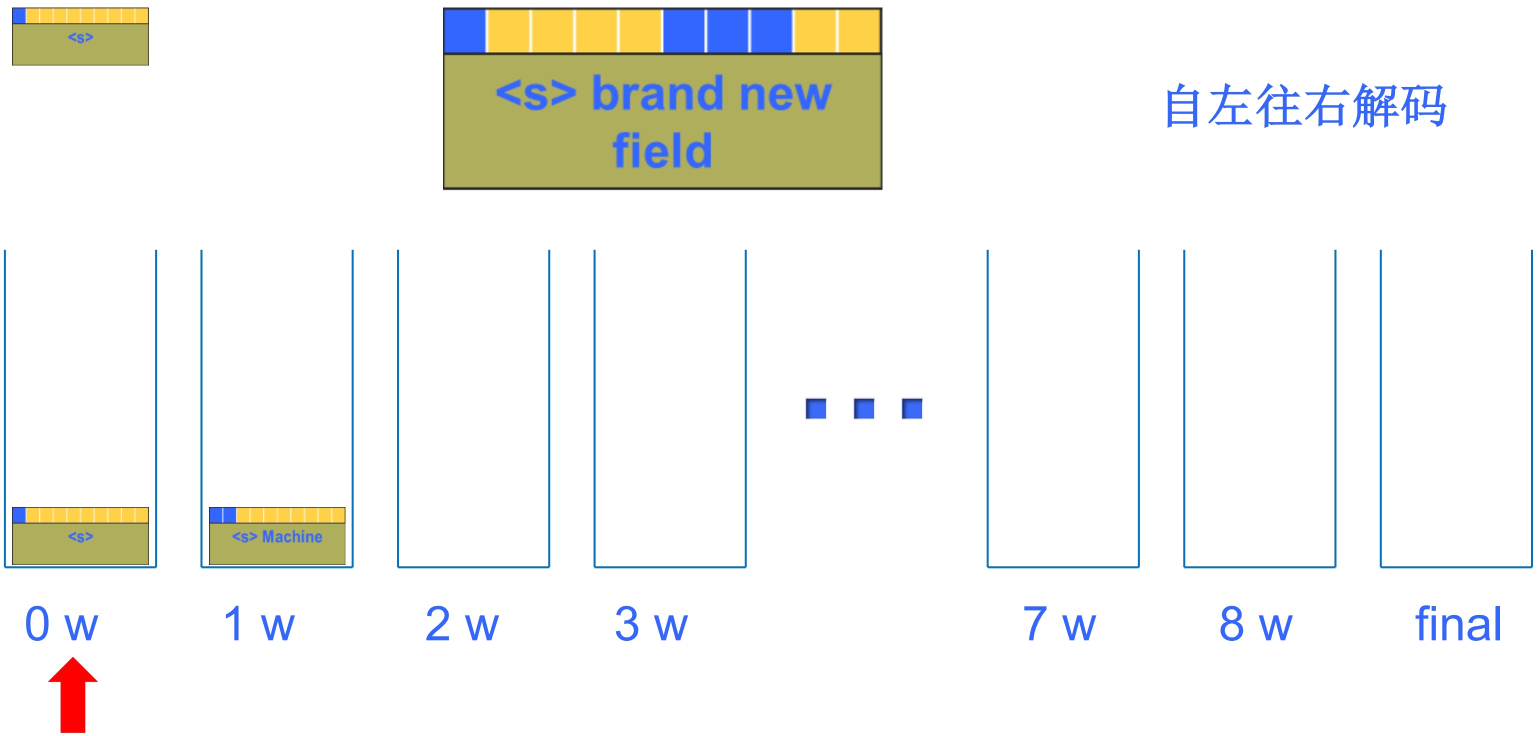


<s> 机器 翻译 是 个 崭新 的 领域 。</s>

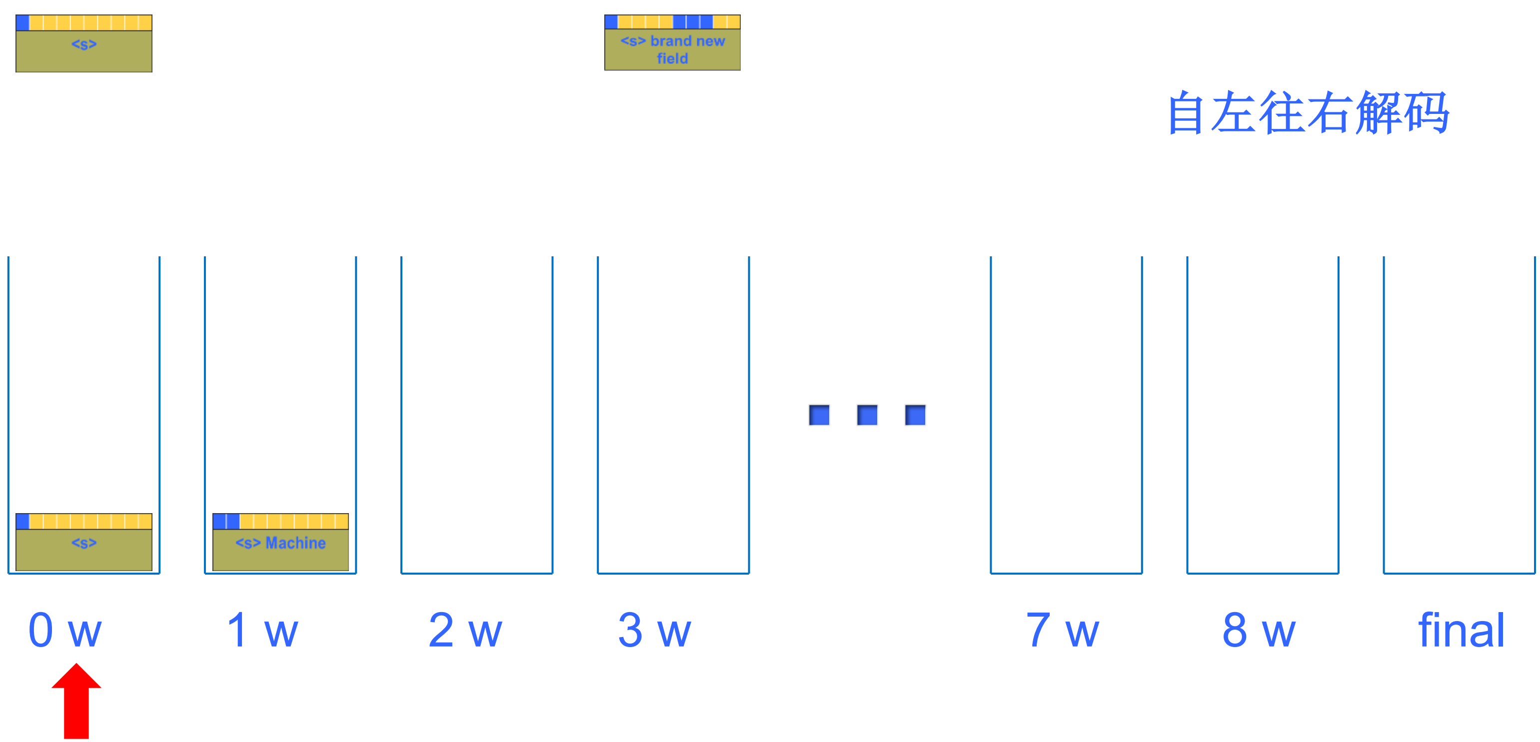
自左往右解码



<s> 机器 翻译 是 个 崭新 的 领域 。</s>

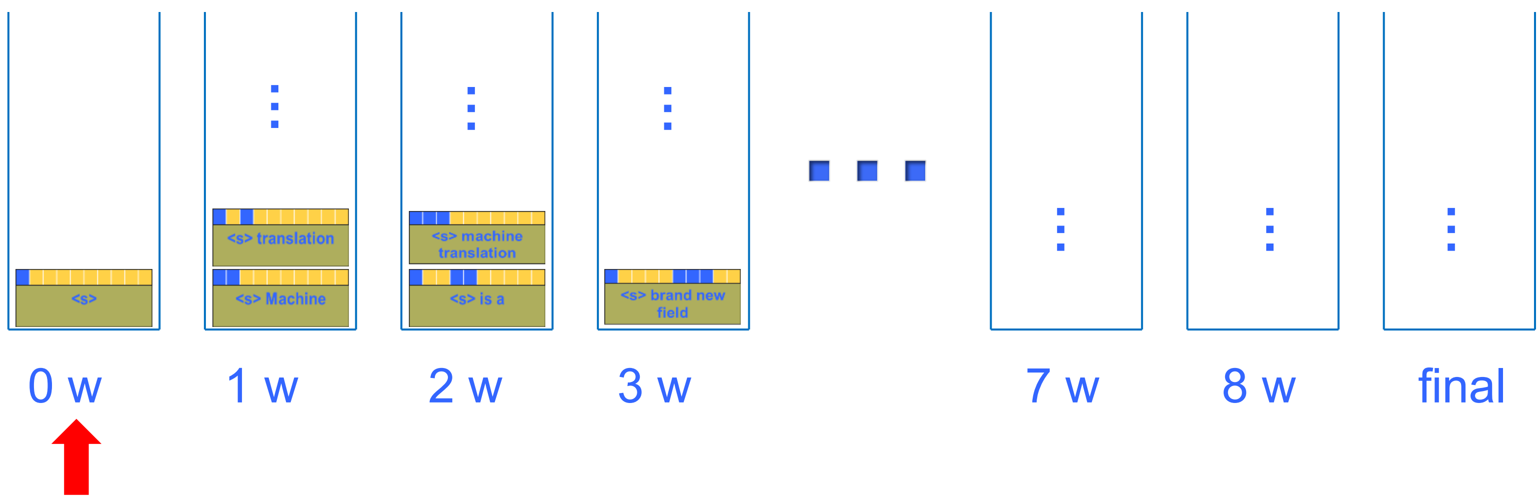


<s> 机器 翻译 是 个 崭新 的 领域 。</s>



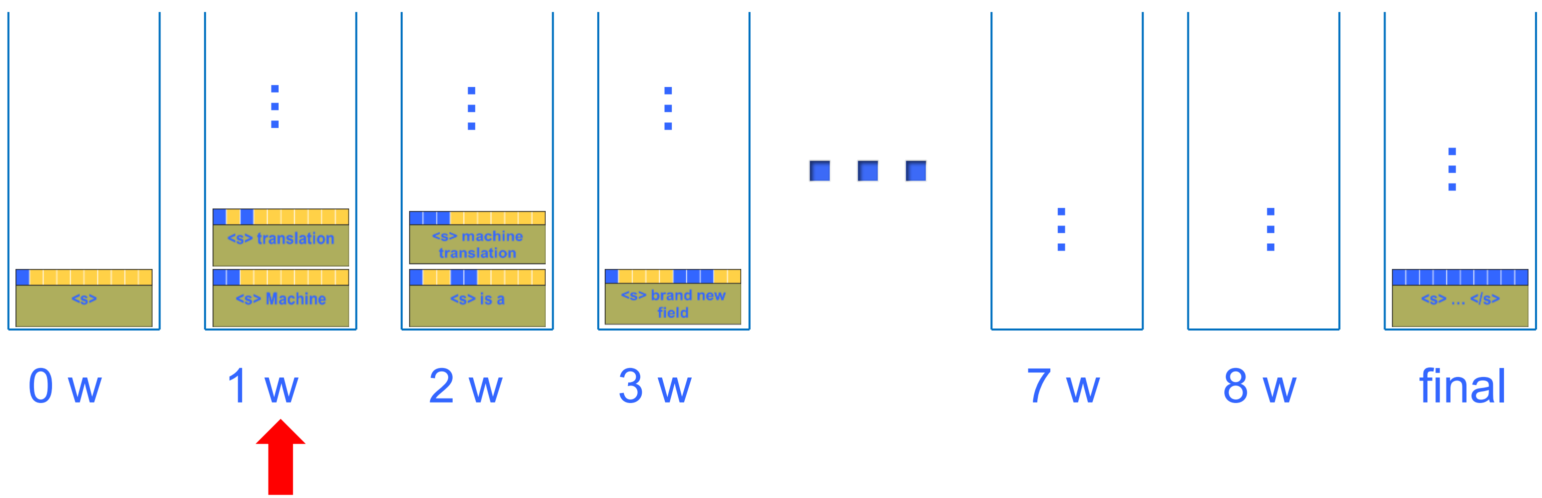
<s> 机器翻译是个崭新的领域。</s>

自左往右解码



<s> 机器 翻译 是 个 崭新 的 领域 。</s>

自左往右解码



机器翻译的人工评价

- 考虑充分性和流畅性

Je suis fatigué.

	Adequacy	Fluency
Tired is I.	5	2
Cookies taste good!	1	5
I am tired.	5	5

难以用于统计系统的自动参数学习

机器翻译的自动评价

- **评价以何为标准?**
  - 人工翻译的结果作为参考译文
  - 使用多个参考译文增强评价结果的鲁棒性
- **如何比较两个句子之间的相似性?**

- WER (Word Error Rate)、PER (Position-Independent WER)
- BLEU (Bilingual Evaluation Understudy)
  - [Papineni et al. 2002]
- TER(Translaiton Error Rate)
  - [Snover et al. 2006]
- Meteor (Metric for Evaluation of Translation with Explicit ORdering)
  - [Lavie and Agarwal 2005]

Word Error Rate (WER)

- 通过给定操作编辑成一致结果的操作数量
  - 编辑距离 (insertion, deletion, substitution)

$$WER = \frac{I + D + S}{N}$$

- 对流畅性把握较好
- 对充分性把握较差
  - 严格匹配

Hypothesis = he saw a man and a woman

Reference = he saw a woman and a man

WER = 2/7

Position-Independent WER (PER)

- **WER对顺序有很强的敏感性，但没有考虑可能发生的整体顺序偏移**
- **PER**
  - 忽略顺序，只考虑单词的匹配(unigram matching)

Hypothesis 1 = he saw a man

Hypothesis 2 = a man saw he

Reference = he saw a man

二者得分相同！

BLEU

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Unigram Precision

- **Unigram Precision of a candidate translation:** $\frac{C}{N}$
  - N is number of words in the candidate,
  - C is the number of words in the candidate which are in at least one reference translation

- **E.g.,**

  Candidate1: It is a guide to action which ensures that the military always obeys the commands of the party.

  precision=17/18

Modified Unigram Precision

- unigram precision存在的问题:

  Candidate: the the the the the the the

  Reference 1: the cat sat on the mat

  Reference 2: there is a cat on the mat

  precision = 7/7 = 100% ???

- 用“Clipping”来进行修正
  - Each word has a “cap”. e.g., cap(the) = 2
  - A candidate word can only be correct a maximum of cap(w) times.
  - cap(w) depends on w’s occurrences in refs.
  - e.g., if cap(the)=2, then precision=2/7

Modified N-Gram Precision

- **容易将modified unigram precision推广到n-gram的情况**
  - 例如：对于之前的示例 candidate 1 和 2 :

$$precision_{bigram}(C1)=10/17$$

$$precision_{bigram}(C2)=1/13$$

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

匹配以外的因素

Candidate 1: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

$$\mathbf{precision_{unigram}(C1)=1}$$

$$\mathbf{precision_{bigram}(C2)=1}$$

准确率 v.s. 召回率？

- 分类中用召回率来评价与准确率相平衡:

  Recall= *C/N*

  C is number of correct n-grams in candidate

  N is number of n-grams in the references.

Candidate 1: I always invariably perpetually do.

Candidate 2: I always do

Reference 1: I always do

Reference 2: I invariably do

Reference 3: I perpetually do

$\text{recall}_{\text{unigram}}(C1)= 5/5$ $\text{recall}_{\text{unigram}}(C2)= 3/5$

Sentence Brevity Penalty

- 取代recall，对过短的句子进行惩罚
- 惩罚标准：
  - 参考译文中最短/最接近的句子
  - e.g. candidate: 12 references: 10 13 15
- 综合不同样本的长度偏好：

$$brevity = \frac{\sum_i r_i}{\sum_i l_i} \qquad BP = \begin{cases} 1 & \text{If } brevity \prec 1 \\ e^{1-brevity} & \text{If } brevity \geq 1 \end{cases}$$

  - e.g. r/l = 1.1, BP = 0.905

BLEU

- 文档级的modified n-gram precision：

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C \in \{Candidate\}} \sum_{ngram \in C} Count(ngram)}$$

- 加上Brevity Penalty

$$Bleu = BP \times (p_1 p_2 p_3 p_4)^{1/4}$$

Translation Error Rate （TER）

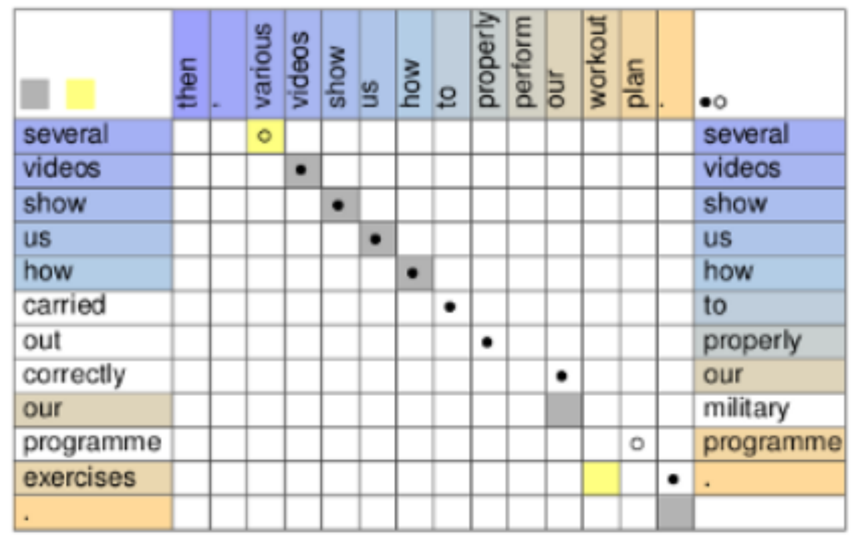
- 如何避免多个参考译文的互相干扰?
  - 选择一个更加匹配的参考译文（edits最少）

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

- Human-targeted TER (HTER)
  - 得到距离最近的“参考译文”
    - 人工编辑翻译结果，直至正确

Meteor (2005-2014)

- 通过更严格的匹配来选择更合适的参考译文，从而提高评价可靠性
  - exact, stem, synonym and paraphrases
  - words/phrases
  - 评价指标的微调
  - 参数权重可调整



Segment 2001

P:	0.650	vs	0.855	:	0.205
R:	0.578	vs	0.689	:	0.111
Frag:	0.522	vs	0.472	:	-0.051
Score:	0.281	vs	0.375	:	0.094

https://www.cs.cmu.edu/~alavie/METEOR/

Reference Graph(Rgraph)

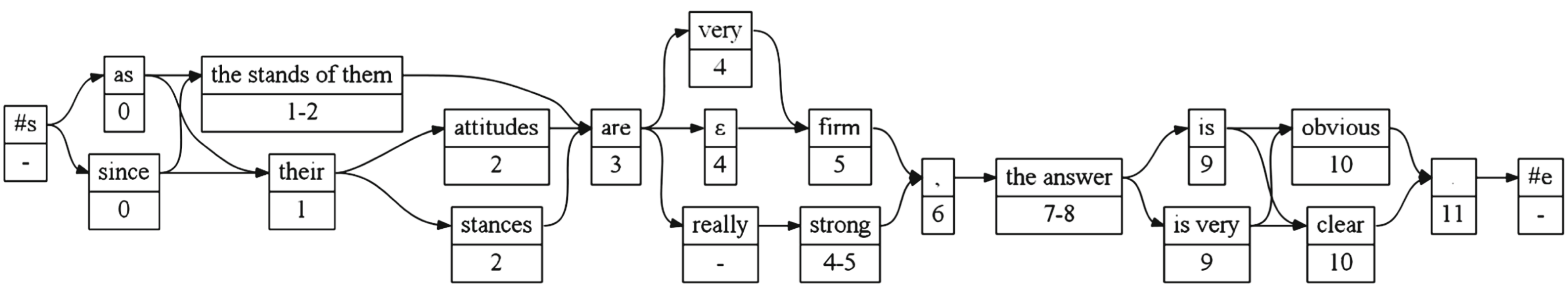
- **构造参考译文图来发掘查找更多可能的翻译**
  - 自动寻找更合适的参考译文

Ref1: as the stands of them are firm , the answer is clear .
Ref2: since their standces are really strong , the answer is very obvious .
Ref3: as their attitudes are very firm , the answer is obvious .
Ref4: since the stands of them are really strong , the answer is obvious .

(a)

Tran1: since their attitudes are firm , the answer is very clear .
Tran2: as since the stands are really very obvious ,

(b)



(c)

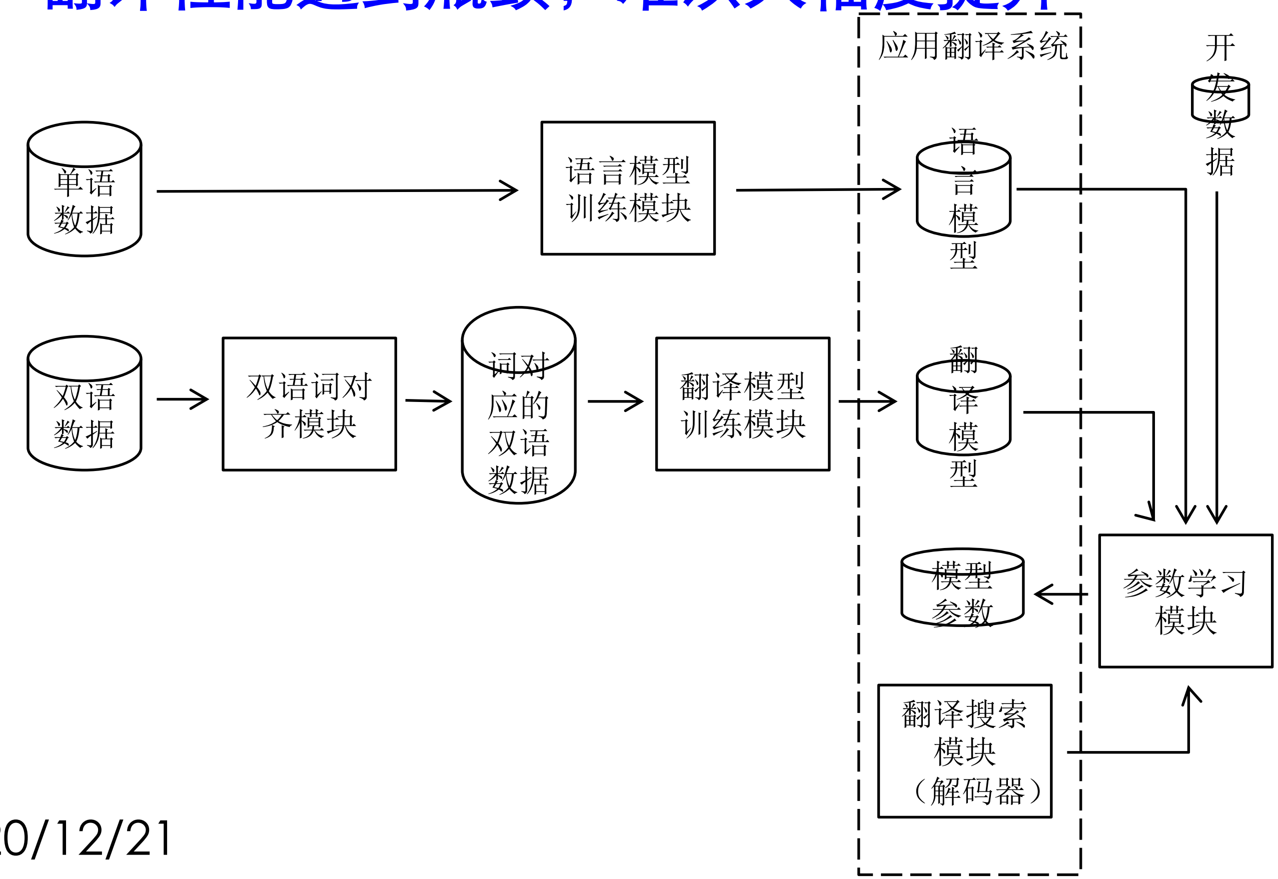
Ji et al. 2017

机器翻译的评估

- **难点**
  - 翻译的多样性
    - 词汇、结构等
  - 意译、省略、语言差别等
- **其他相关研究**
  - 评价文档相关的翻译质量
    - 翻译词汇一致性
    - 时态、语态等状态一致性
    - 话题转换、语言流畅程度等
- **神经网络可能能带来更大的突破**

统计机器翻译回顾

- 可以一定程度上从数据中自动挖掘翻译知识
- 流程相对复杂，其中各个部分都不断被改进和优化
- 翻译性能遇到瓶颈，难以大幅度提升



ACL (Natural Language Processing)		
2018	Finding syntax in human encephalography with beam search	John Hale, Cornell University; et al.
2017	Probabilistic Typology: Deep Generative Models of Vowel Inventories	Ryan Cotterell & Jason Eisner, Johns Hopkins University
2016	Finding Non–Arbitrary Form–Meaning Systematicity Using String–Metric Learning for Kernel Regression	E. Darío Gutiérrez, University of California Berkeley; et al.
2015	Improving Evaluation of Machine Translation Quality Estimation	Yvette Graham, Trinity College Dublin
	Learning Dynamic Feature Selection for Fast Sequential Prediction	Emma Strubell, University of Massachusetts Amherst; et al.
2014	Fast and Robust Neural Network Joint Models for Statistical Machine Translation	Jacob Devlin, Raytheon BBN Technologies; et al.
2013	Grounded Language Learning from Video Described with Sentences	Haonan Yu & Jeffrey Mark Siskind, Purdue University
2012	String Re–writing Kernel	Fan Bu, Tsinghua University; et al.
	Bayesian Symbol–Refined Tree Substitution Grammars for Syntactic Parsing	Hiroyuki Shindo, NTT Communication Science Laboratories; et al.
2011	Unsupervised Part–of–Speech Tagging with Bilingual Graph–Based Projections	Dipanjan Das, Carnegie Mellon University Slav Petrov, Google
2010	Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates	Matthew Gerber & Joyce Y. Chai, Michigan State University
2009	Reinforcement Learning for Mapping Instructions to Actions	S.R.K. Branavan, Massachusetts Institute of Technology; et al.
	K–Best A* Parsing	Adam Pauls & Dan Klein, University of California Berkeley
	Concise Integer Linear Programming Formulations for Dependency Parsing	André F.T. Martins, Instituto de Telecomunicações; et al.
2008	Forest Reranking: Discriminative Parsing with Non–Local Features	Liang Huang, University of Pennsylvania
	A New String–to–Dependency Machine Translation Algorithm with a Target Dependency Language Model	Libin Shen, BBN Technologies; et al.
2007	Learning synchronous grammars for semantic parsing with lambda calculus	Yuk Wah Wong & Raymond J. Mooney, University of Texas at Austin
2006	Semantic taxonomy induction from heterogenous evidence	Rion Snow, Stanford University; et al.
2005	A Hierarchical Phrase–Based Model for Statistical Machine Translation	David Chiang, University of Maryland
2004	Finding Predominant Word Senses in Untagged Text	Diana McCarthy, University of Sussex; et al.
2003	Accurate Unlexicalized Parsing	Dan Klein & Christopher D. Manning, Stanford University
	Towards a Model of Face–to–Face Grounding	Yukiko I. Nakano, RISTEX; et al.
2002	Discriminative Training and Maximum Entropy Models for Statistical Machine Translation	Franz Josef Och & Hermann Ney, RWTH Aachen University
2001	Immediate–Head Parsing for Language Models	Eugene Charniak, Brown University
	Fast Decoding and Optimal Decoding for Machine Translation	Ulrich Germann, University of Southern California; et al.

参考文献

- Warren Weaver. Translation. 1949
- Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers. 1984
- P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematic of statistical machine translation: Parameter estimation. Computational Linguistics 1993
- P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in HLT-NAACL, 2003.
- D. Chiang, “A hierarchical phrase-based model for statistical machine translation,” in annual meeting of the Association for Computational Linguistics, 2005.
- K. Yamada and K. Knight, “A syntax-based statistical translation model,” in Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics, 2001, pp. 523–530.

参考文献

- Y. Liu, Q. Liu, and S. Lin, “Tree-to-string alignment template for statistical machine trans- lation,” in Proceedings of the 44th Annual Meeting of the Association of Computational Linguistics. The Association for Computer Linguistics, 2006.
- F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation,” in Proceedings of the 40th Annual Meeting on Association for Compu- tational Linguistics, ser. ACL ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 295–302.
- F. J. Och, “Minimum error rate training in statistical machine translation,” in ACL ’03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Mor- ristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 160–167.
- David Chiang, Kevin Knight, and Wei Wang. 11,001 new features for statistical machine translation. In Proc. NAACL HLT, 218–226. 2009. Best paper award.
- M. Snover, B. J. Dorr, and R. Schwartz, “A study of translation edit rate with targeted human annotation,” in Proceedings of AMTA, 2006.

参考文献

- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in ACL ’02: Proceedings of the 40th Annual Meeting on Associa- tion for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 2002, pp. 311–318.
- A. Lavie and A. Agarwal, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005, pp. 65–72.
- Hongjie Ji, Shujian Huang*, Qi Hou, Cunyan Yin, and Jiajun Chen. Rgraph:Generating reference graphs for better machine translation evaluation. In China Workshop on Machine Translation, pages 55-67. Springer, 2017.