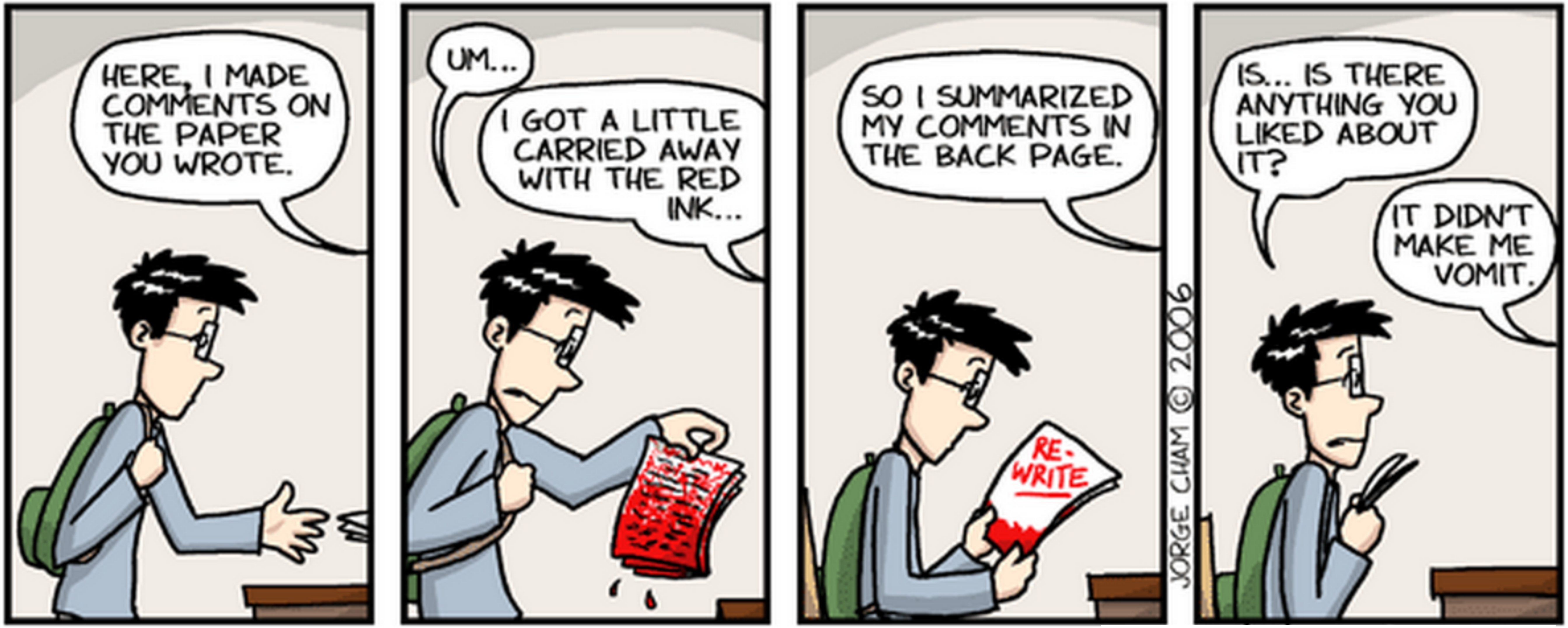
如何撰写高质量科技论文

刘洋

HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.

HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.
UM...
I GOT A LITTLE CARRIED AWAY WITH THE RED INK...

HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.
UM...
I GOT A LITTLE CARRIED AWAY WITH THE RED INK...
SO I SUMMARIZED MY COMMENTS IN THE BACK PAGE.
RE-WRITE


HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.
UM...
I GOT A LITTLE CARRIED AWAY WITH THE RED INK...
SO I SUMMARIZED MY COMMENTS IN THE BACK PAGE.
RE-WRITE
JORGE CHAM © 2006
IS... IS THERE ANYTHING YOU LIKED ABOUT IT?
IT DIDN'T MAKE ME VOMIT.

HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.
UM...
I GOT A LITTLE CARRIED AWAY WITH THE RED INK...
SO I SUMMARIZED MY COMMENTS IN THE BACK PAGE.
RE-WRITE
JORGE CHAM © 2006
IS... IS THERE ANYTHING YOU LIKED ABOUT IT?
IT DIDN'T MAKE ME VOMIT.
www.phdcomics.com

HERE, I MADE COMMENTS ON THE PAPER YOU WROTE.
UM...
I GOT A LITTLE CARRIED AWAY WITH THE RED INK...
SO I SUMMARIZED MY COMMENTS IN THE BACK PAGE.
RE-WRITE
IS... IS THERE ANYTHING YOU LIKED ABOUT IT?
IT DIDN'T MAKE ME VOMIT.
JORGE CHAM © 2006
www.phdcomics.com

如何写出高质量的学术论文?

论文发表流程

发表论文

论文发表流程

发表论文

确定方向

论文发表流程

发表论文

确定问题

确定方向

论文发表流程

发表论文

确定思路

确定问题

确定方向

论文发表流程

确定方向

确定问题

确定思路

确定方法

发表论文

论文发表流程

发表论文

实验验证

确定方法

确定思路

确定问题

确定方向

论文发表流程

确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

发表论文

论文发表流程

确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

发表论文

写论文时什么最重要?

确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

写论文时什么最重要?

确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

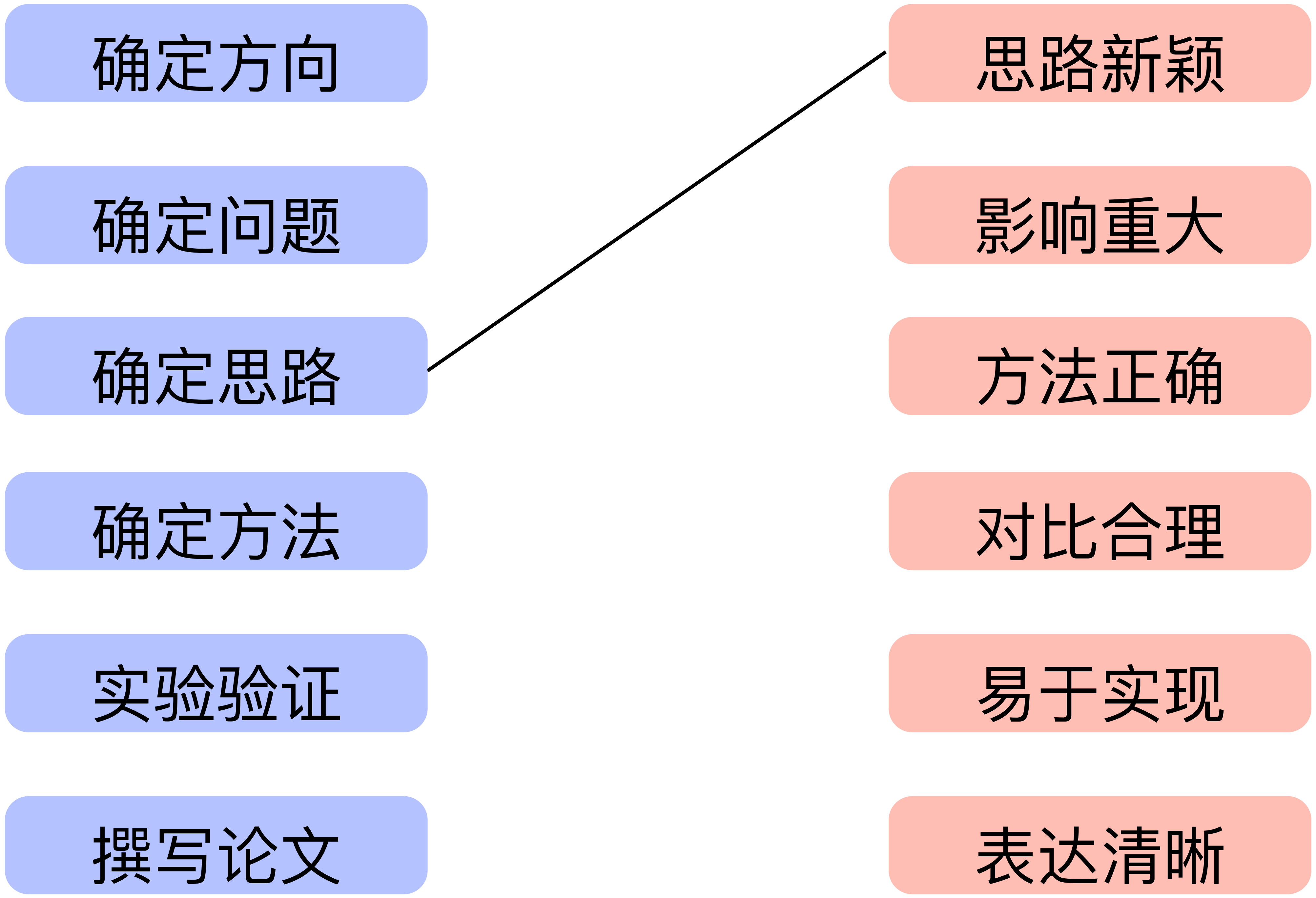
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

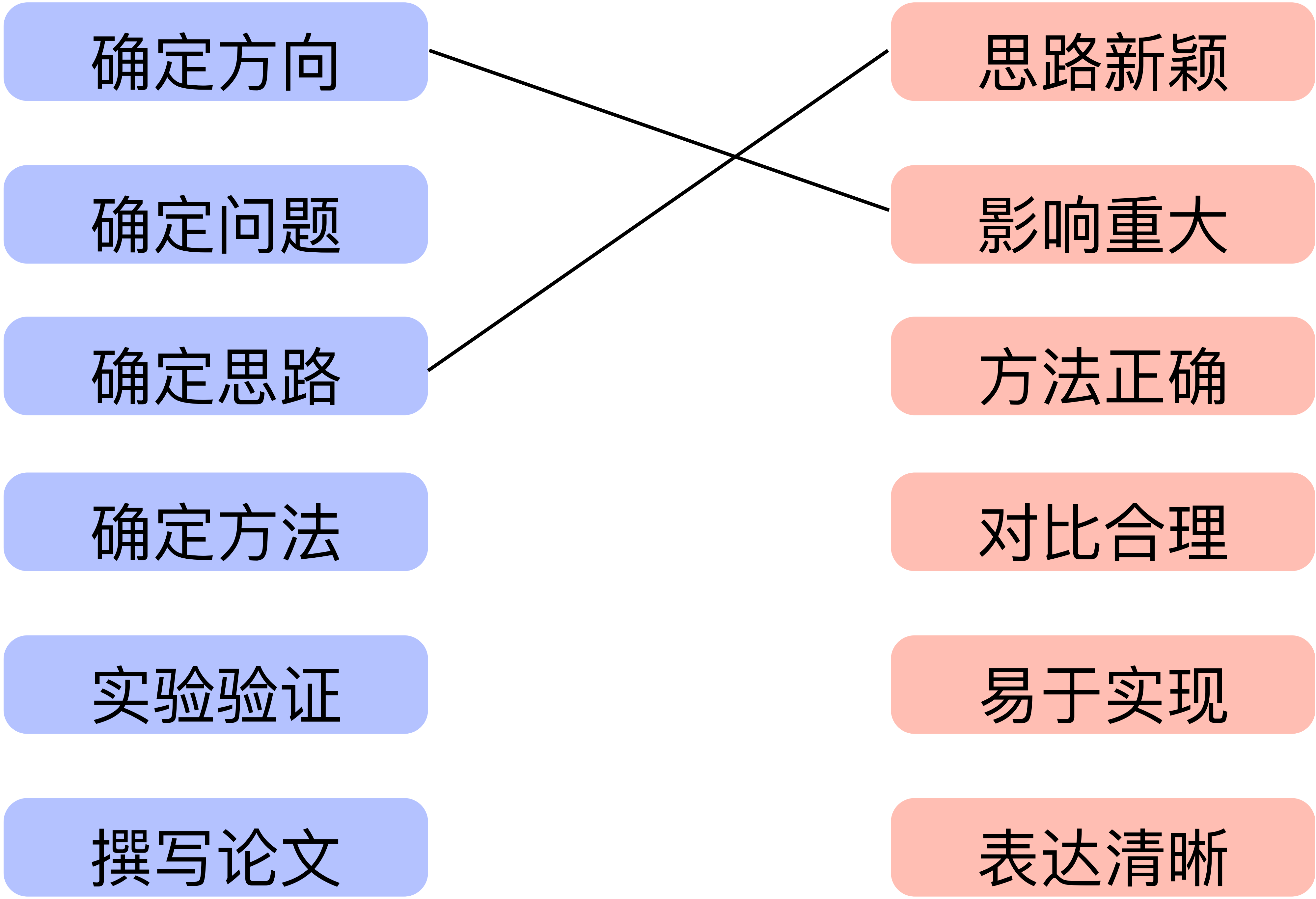
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

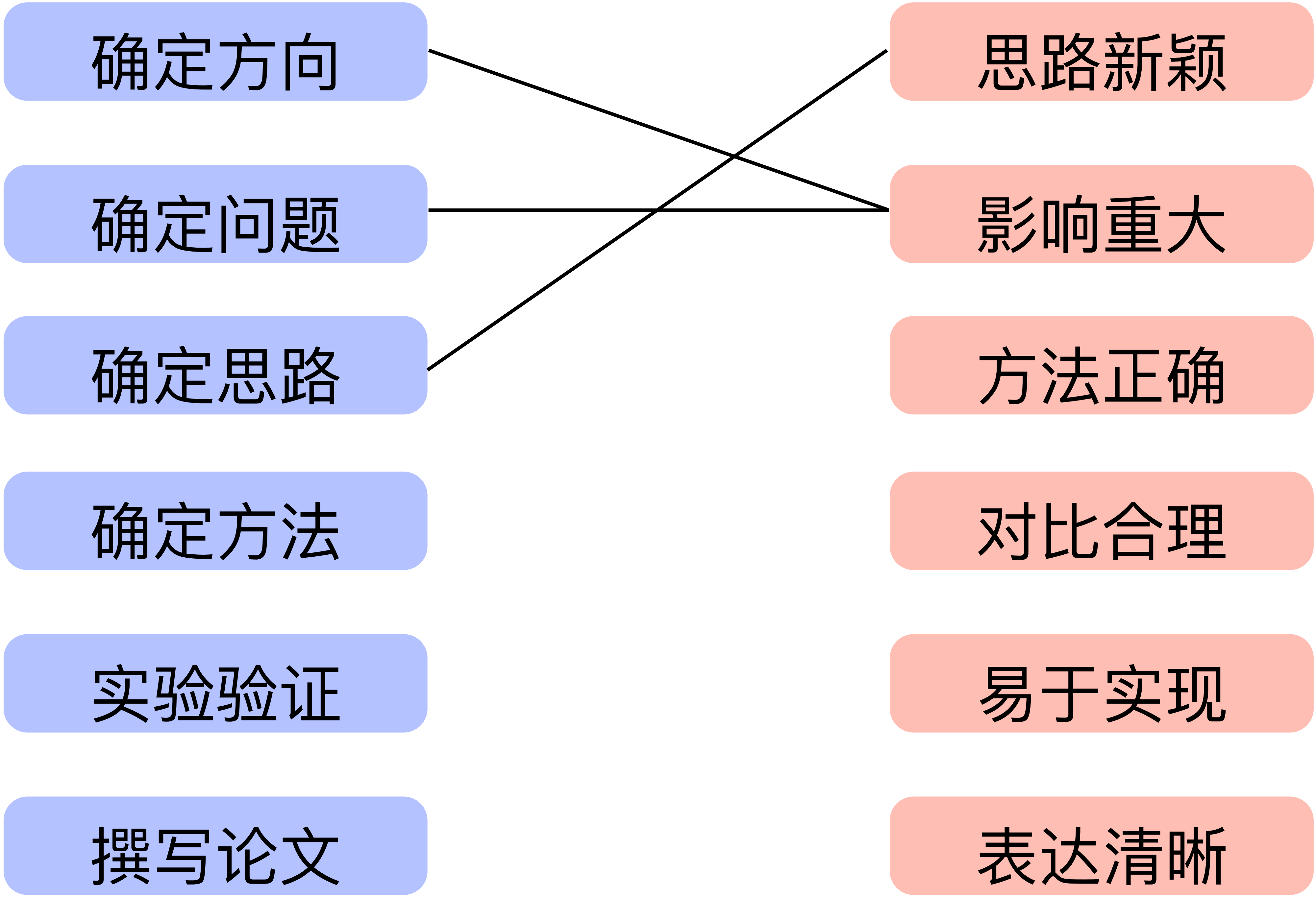
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

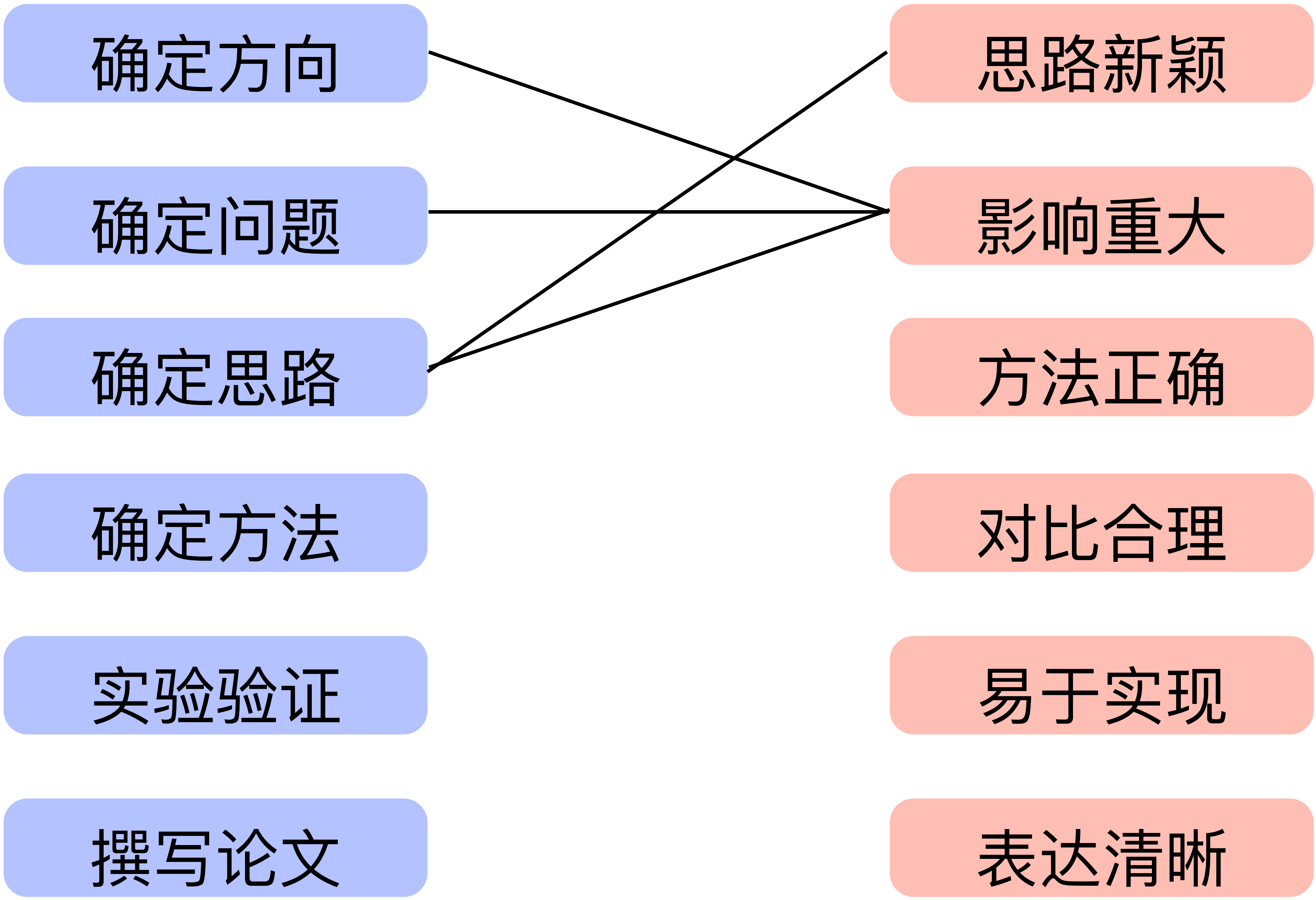
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

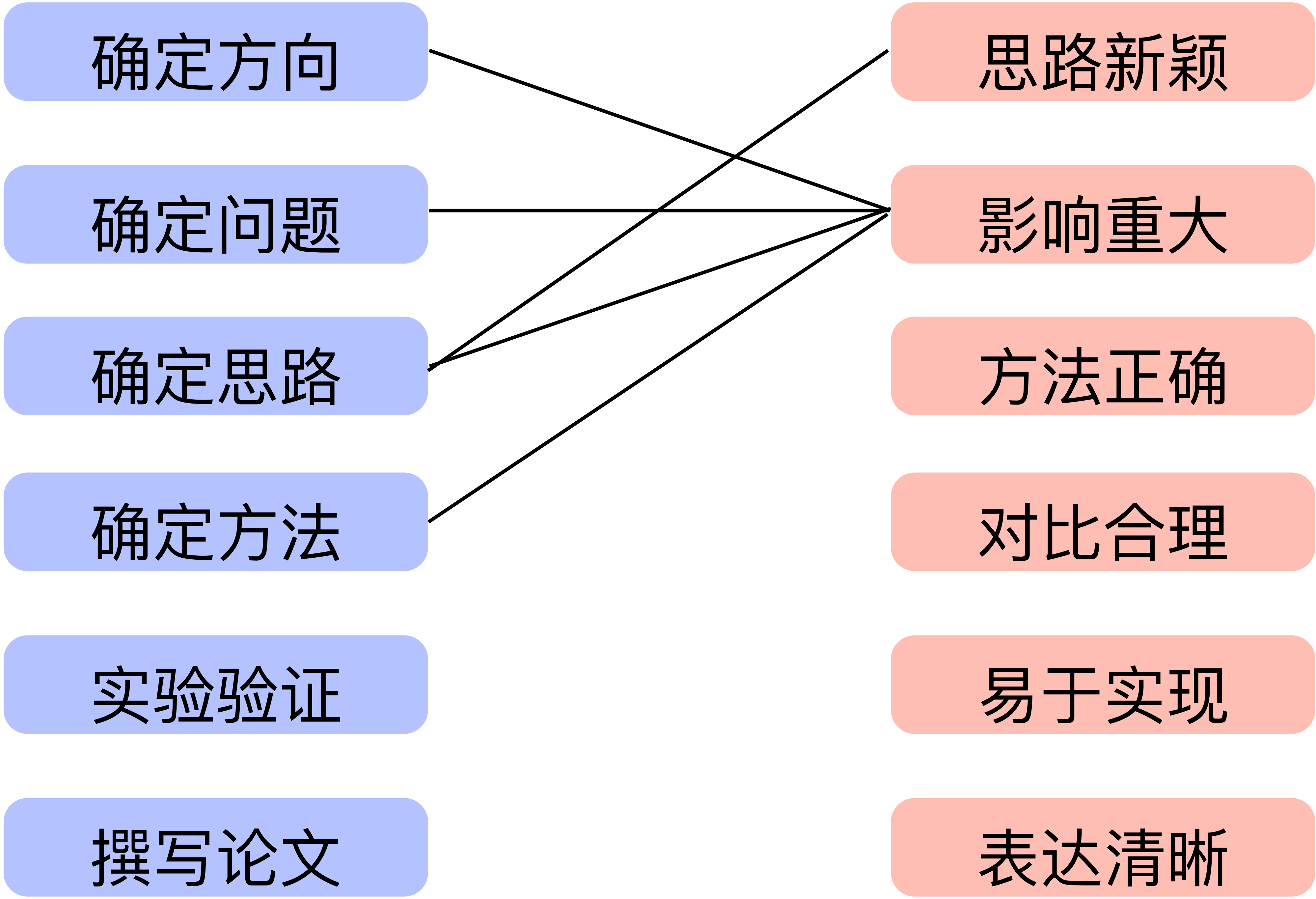
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

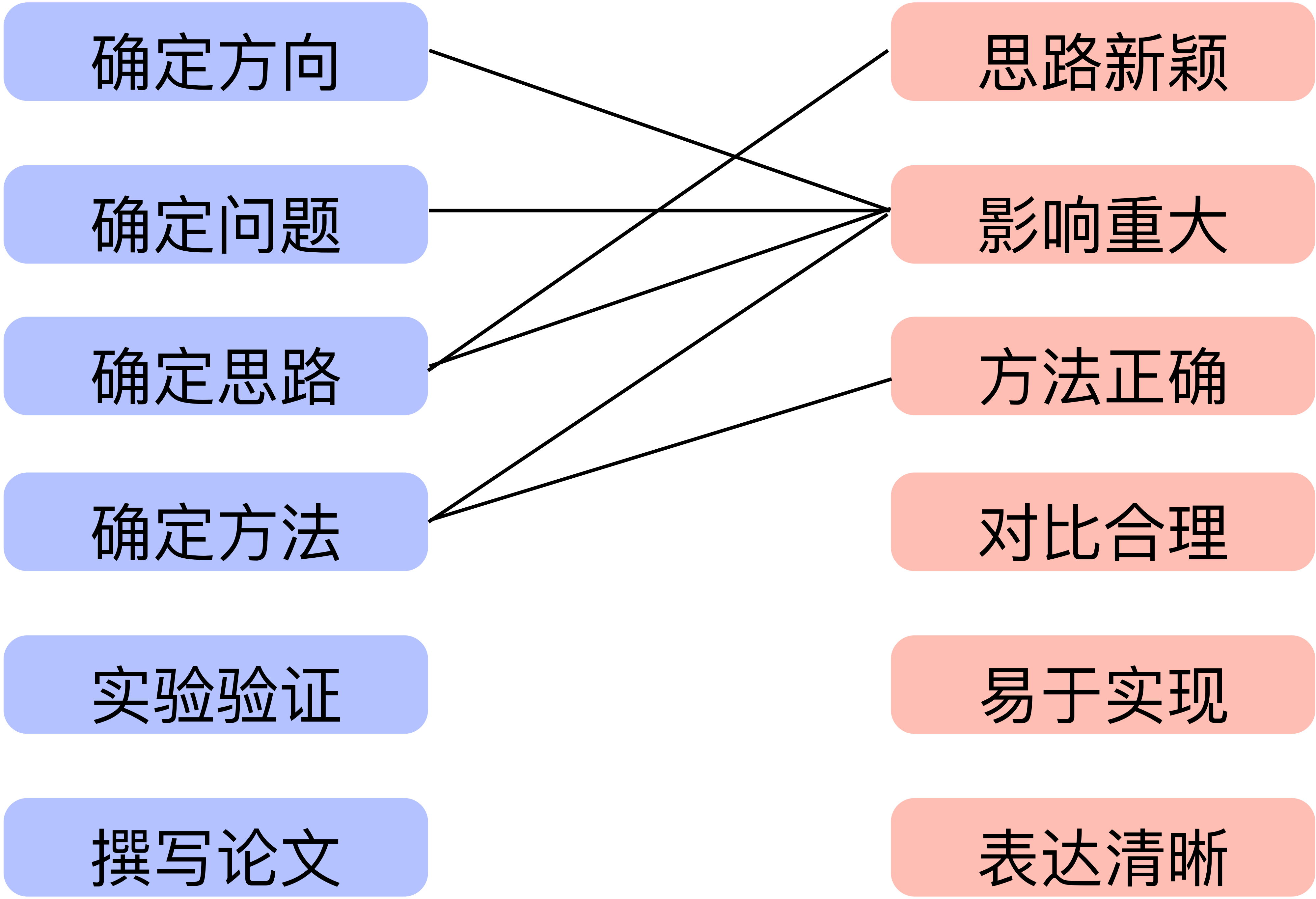
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

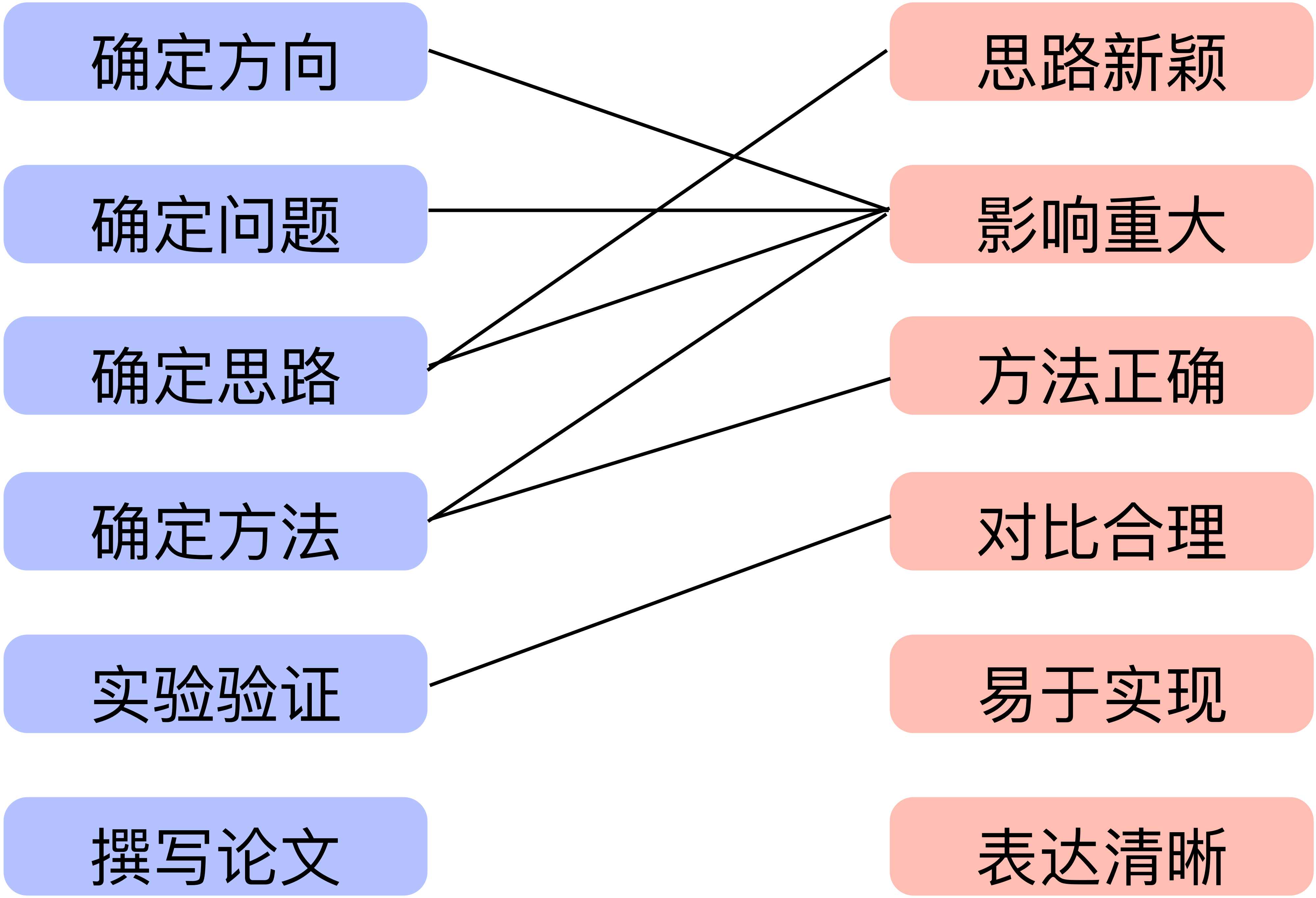
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

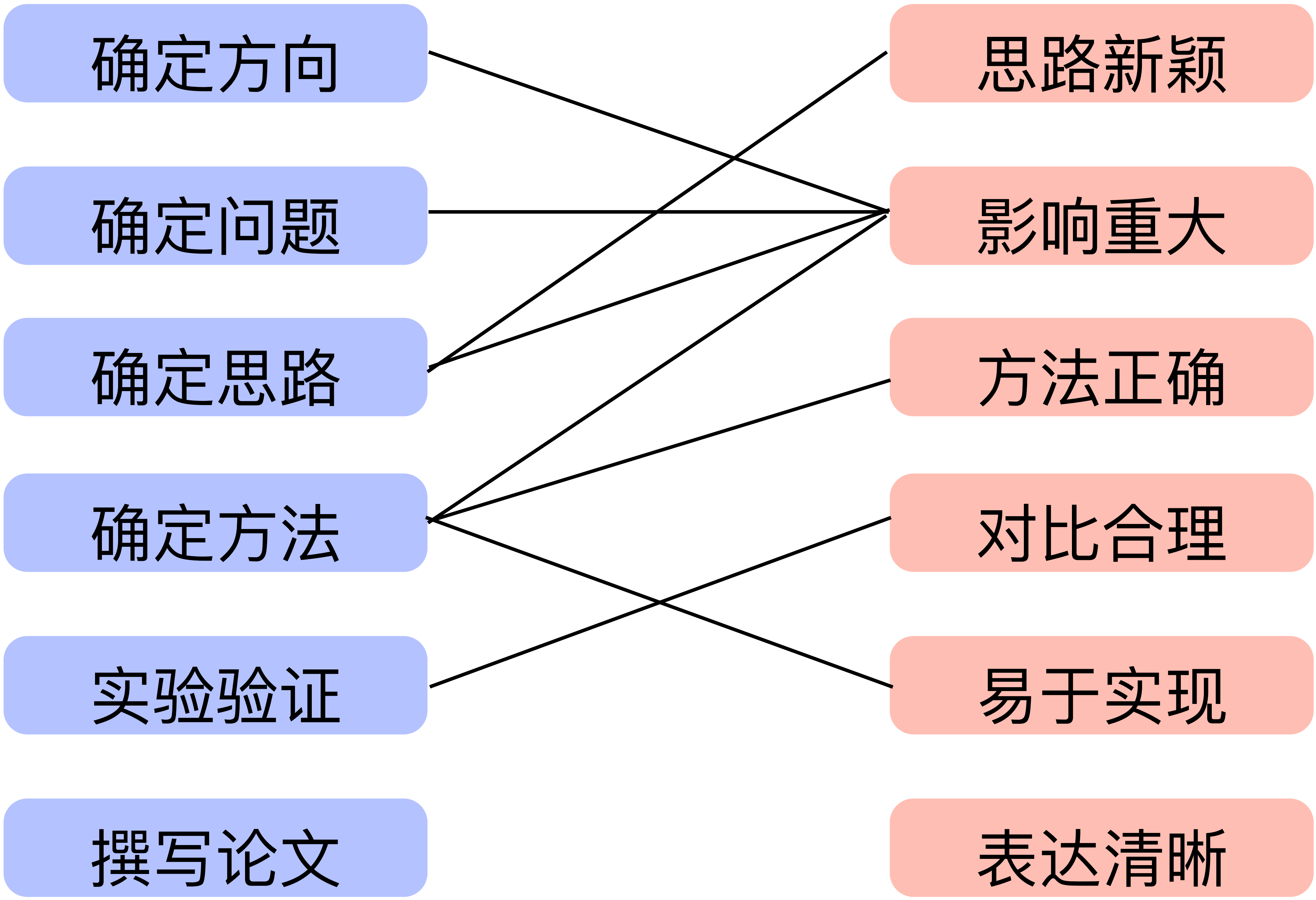
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

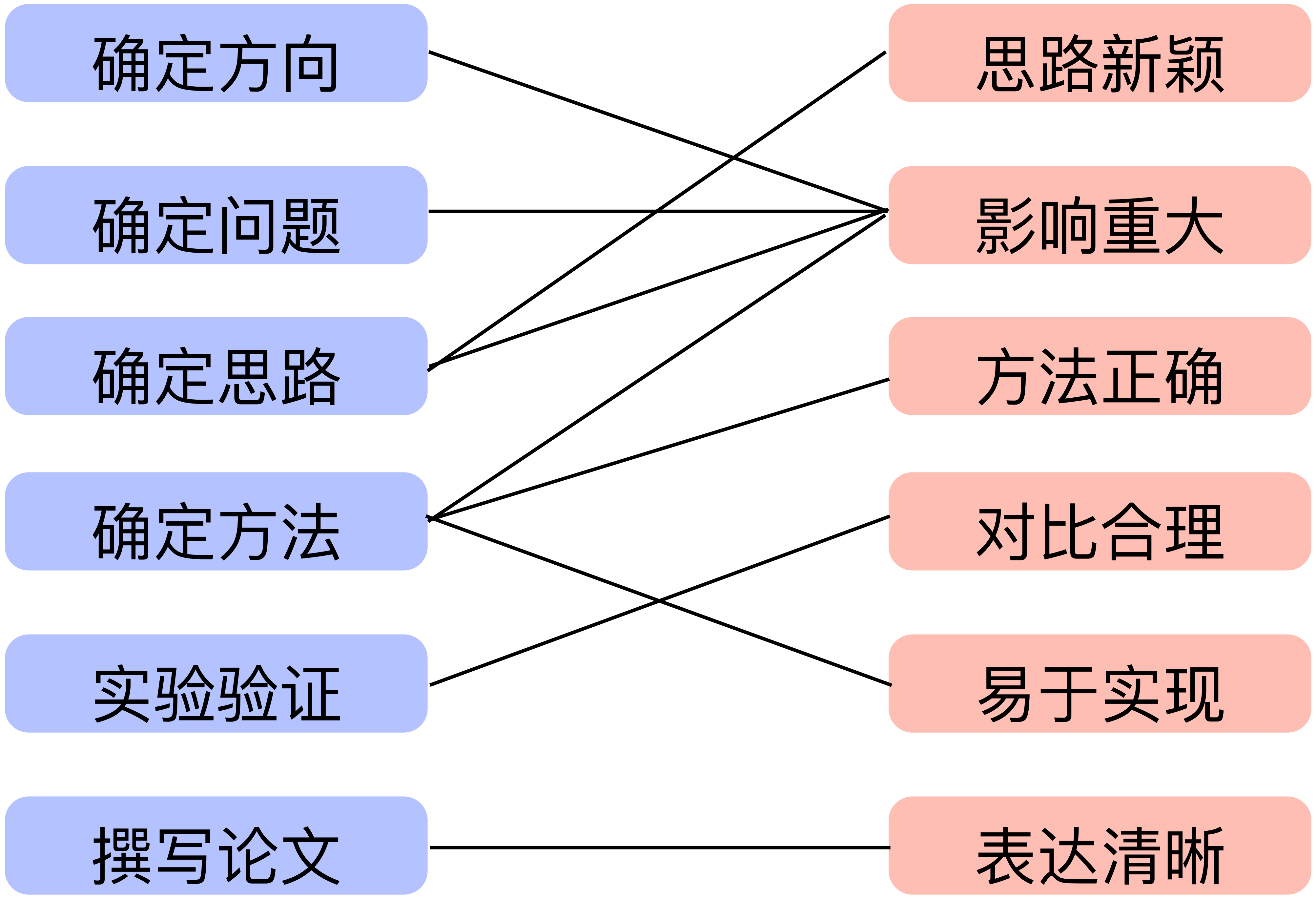
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

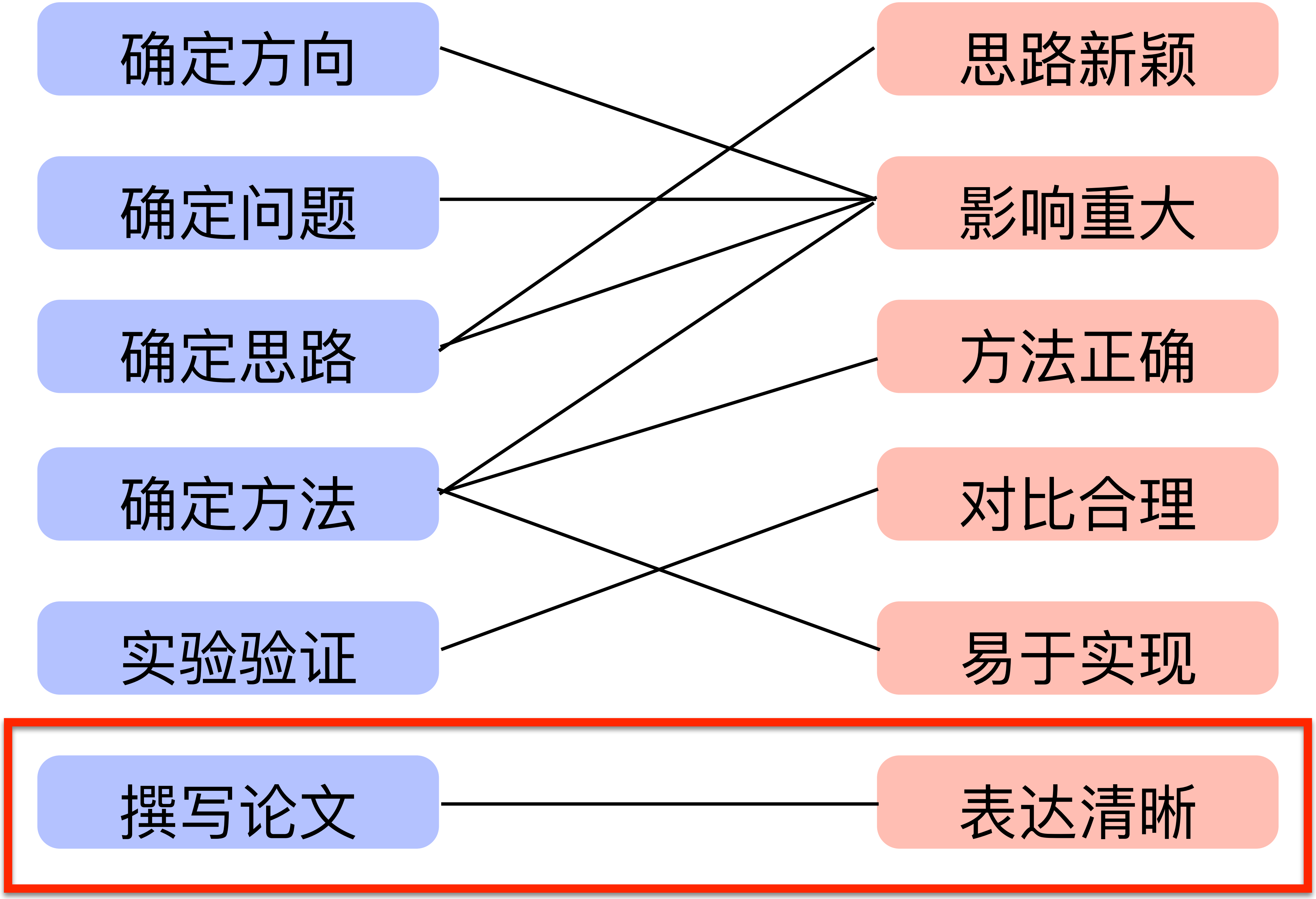
方法正确

对比合理

易于实现

表达清晰

写论文时什么最重要?



确定方向

确定问题

确定思路

确定方法

实验验证

撰写论文

思路新颖

影响重大

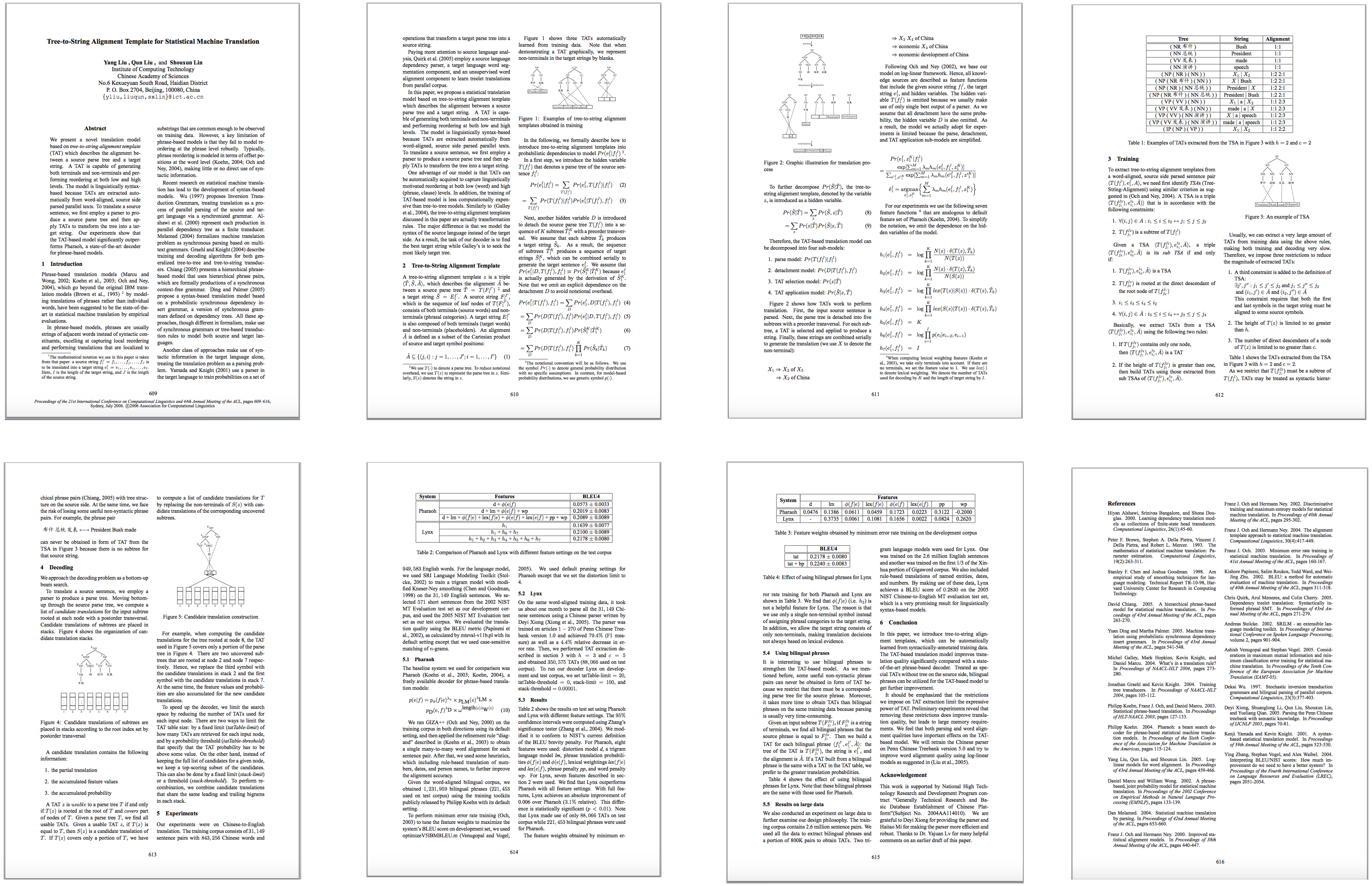
方法正确

对比合理

易于实现

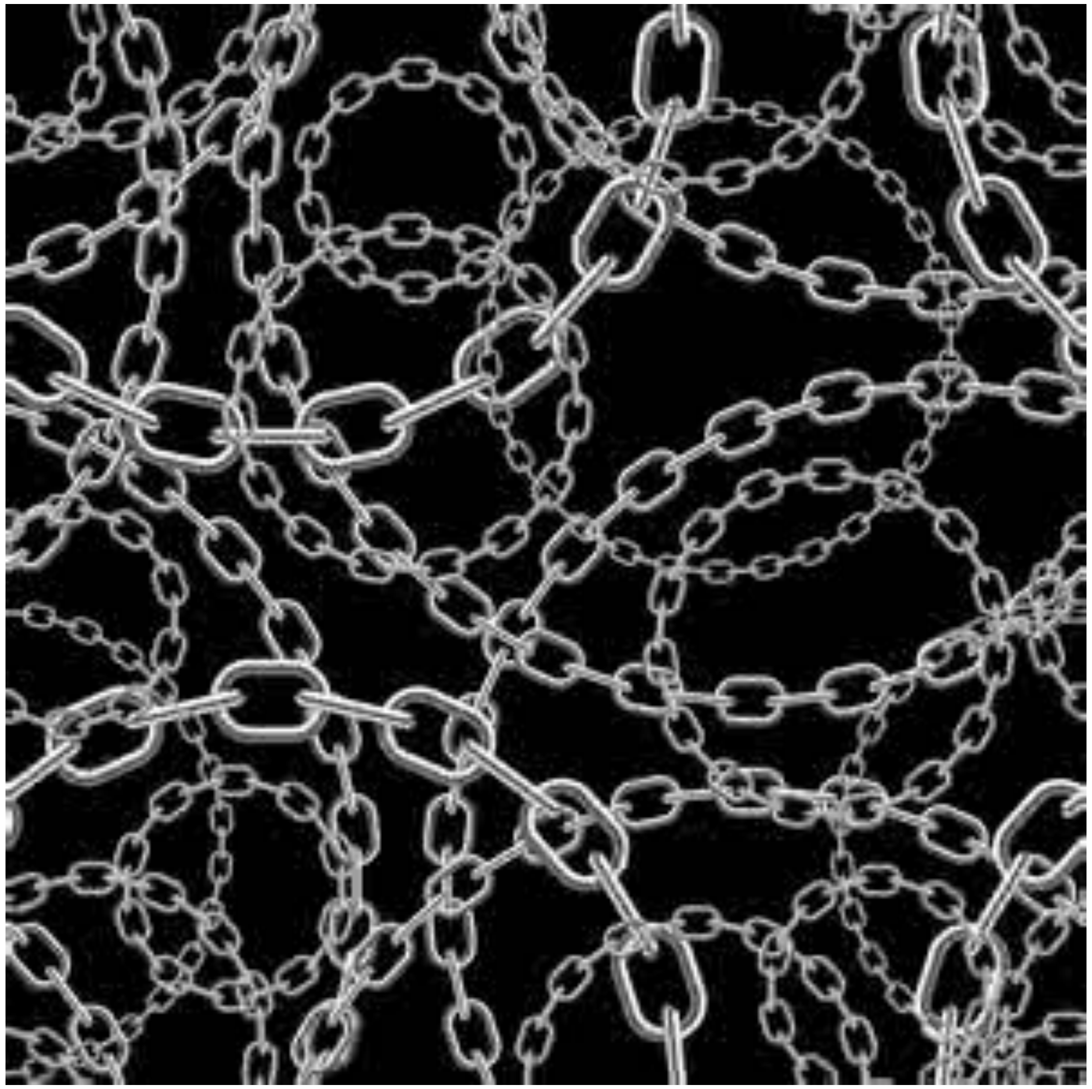
表达清晰

信息



读者看到的是论文

逻辑

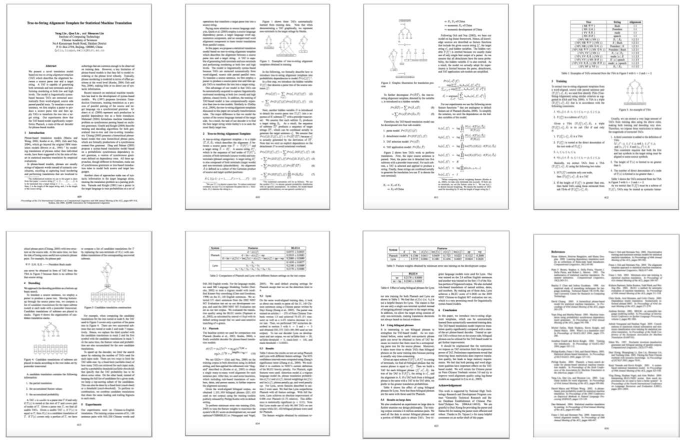


信息的传递按逻辑来组织

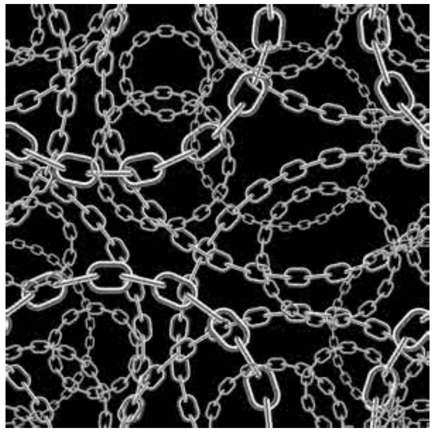
思想

深层次反映的是作者的思想

信息、逻辑与思想



信息为**表**

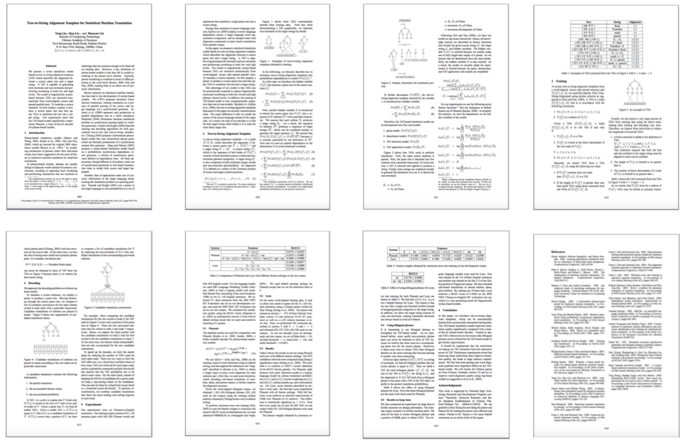


逻辑为**骨**

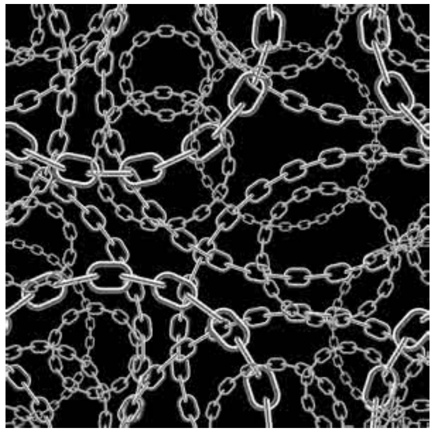
思想为**心**

阅读与写作

信息



逻辑

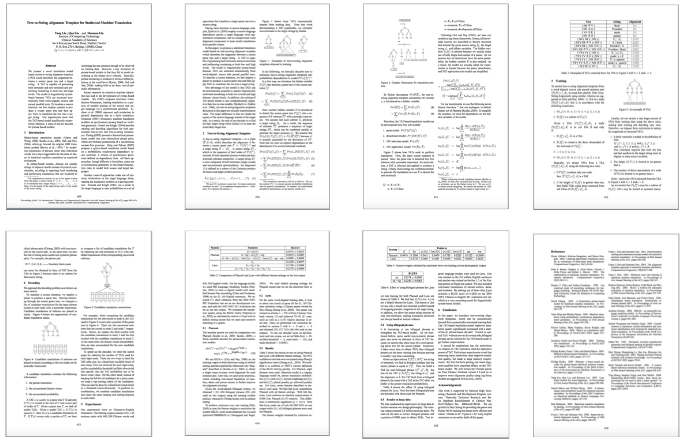


思想

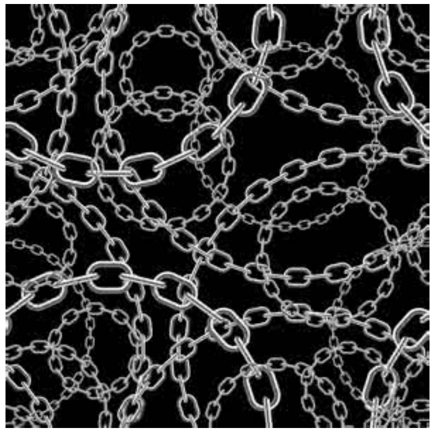
阅读与写作

阅读

信息



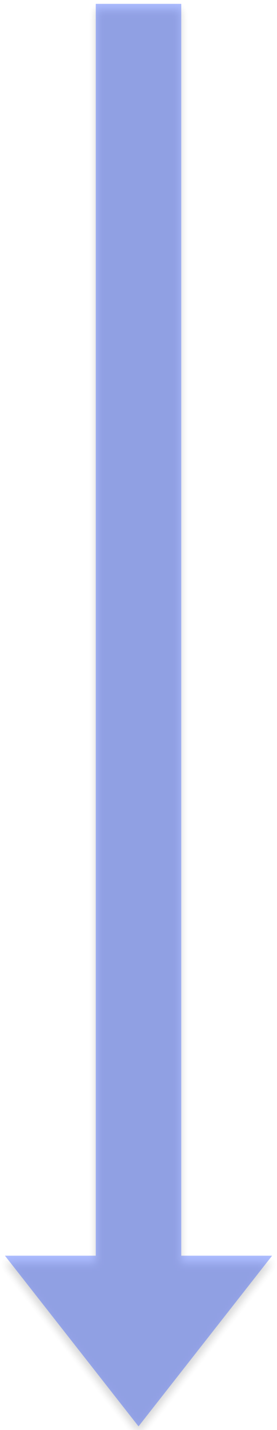
逻辑



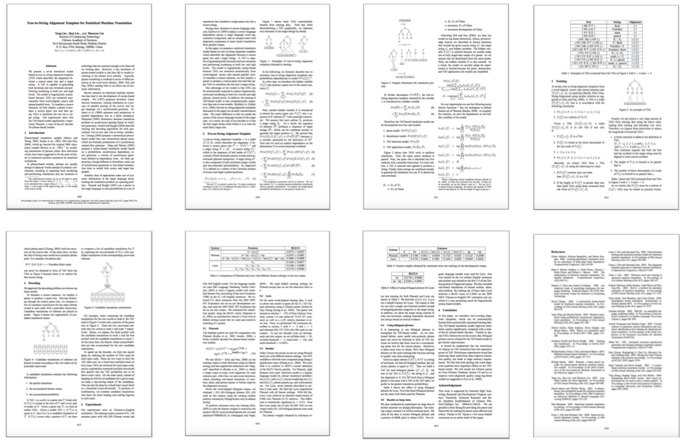
思想

阅读与写作

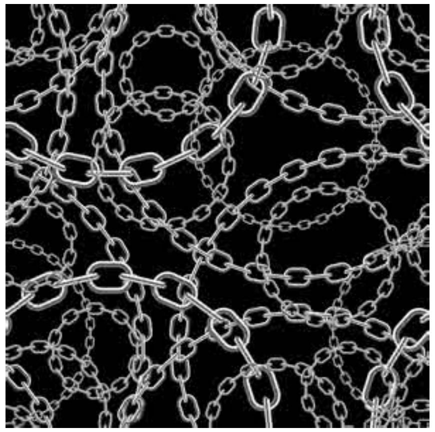
阅读



信息



逻辑

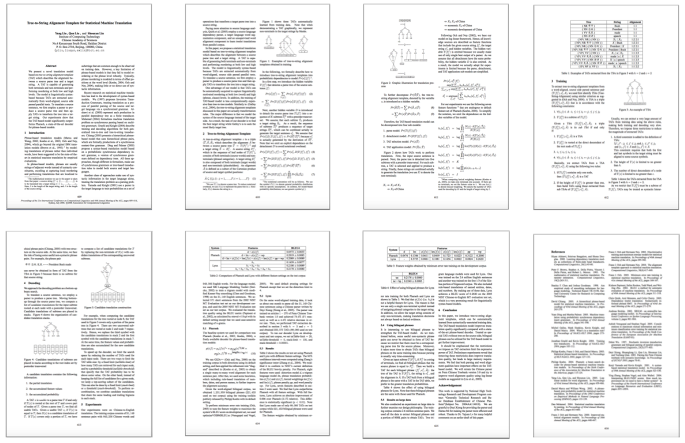


思想

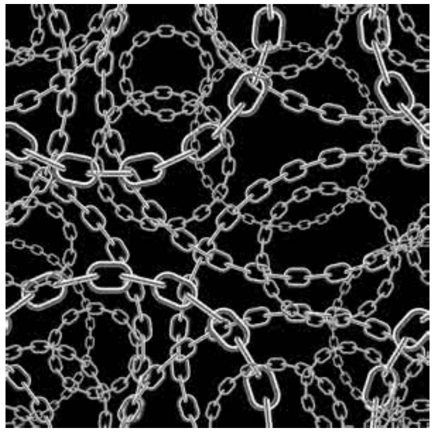
阅读与写作

阅读 写作

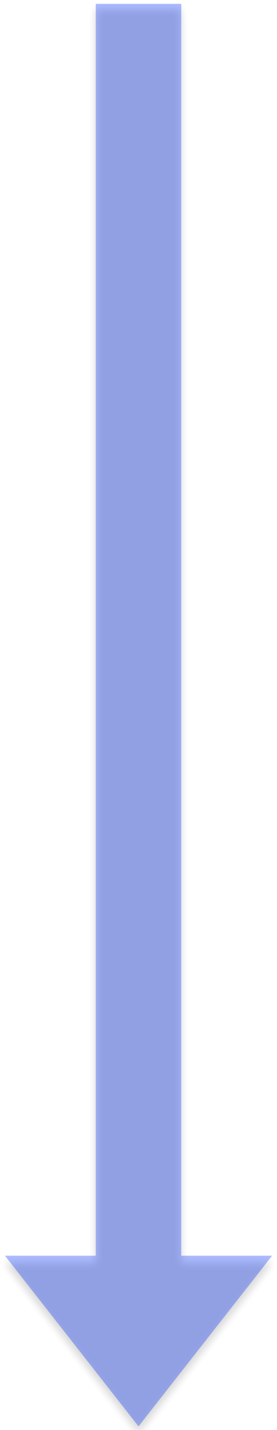
信息



逻辑

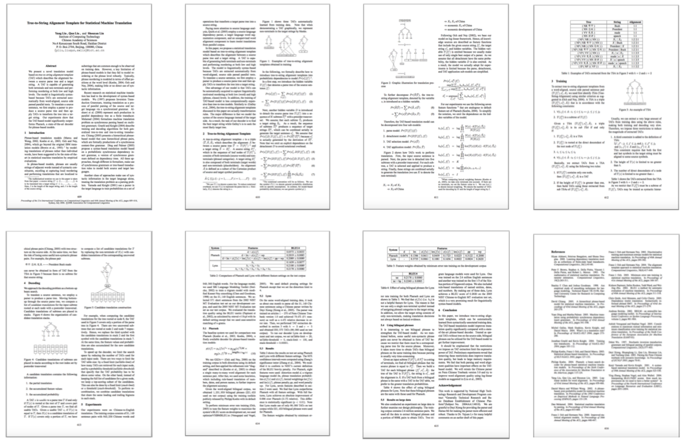


思想

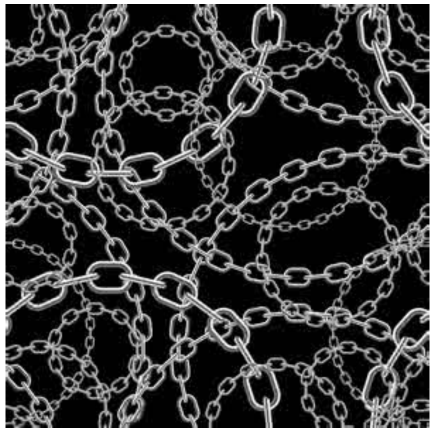


阅读与写作

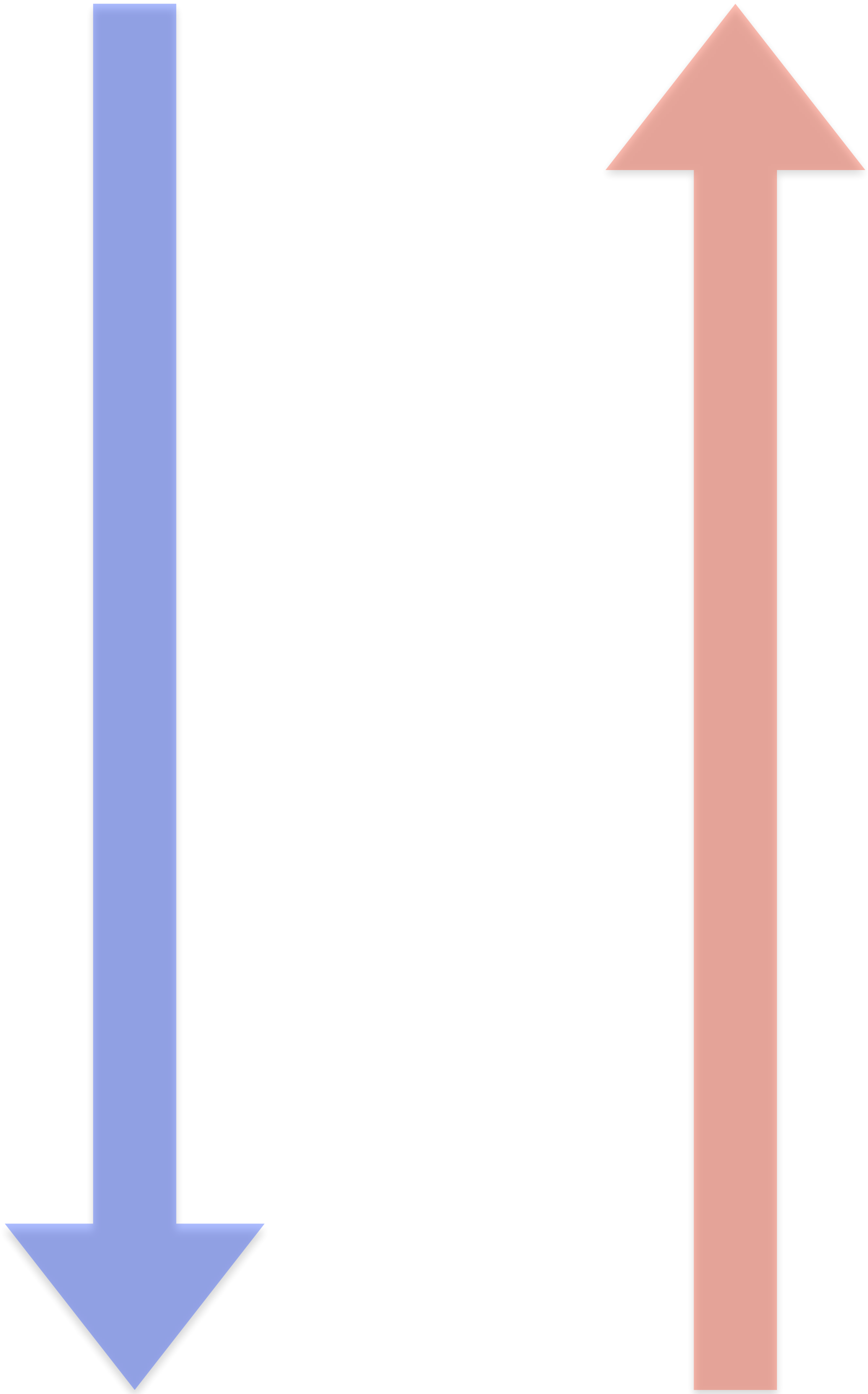
信息



逻辑



思想



阅读

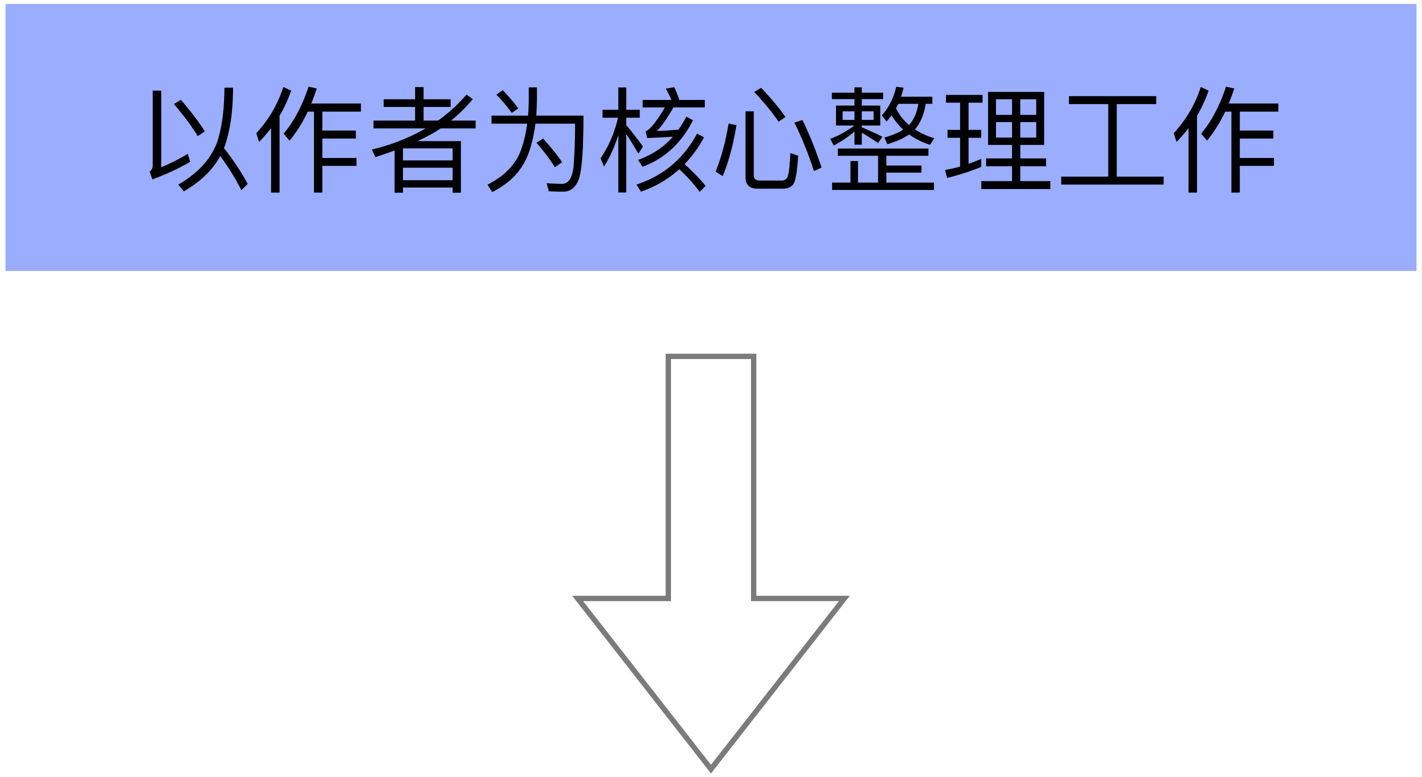
写作

观念转变

观念转变

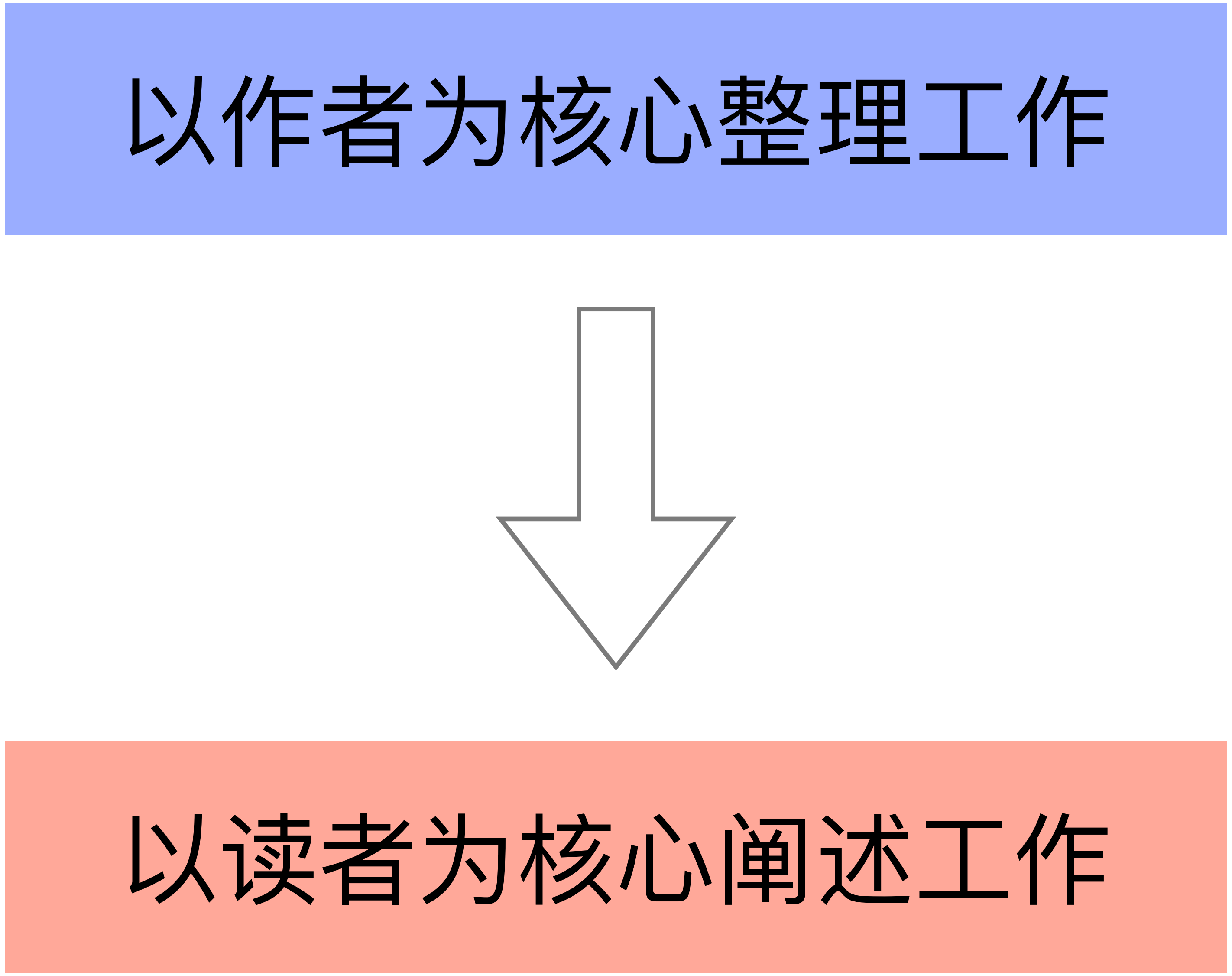
以作者为核心整理工作

观念转变



以作者为核心整理工作

观念转变



以作者为核心整理工作

以读者为核心阐述工作

全心全意为读者服务

全心全意为读者服务

信息的呈现符合读者的认知惯性

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

尽量降低读者的理解难度

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

尽量降低读者的理解难度

合理地综合使用信息元素：图>曲线>表>正文>公式

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

尽量降低读者的理解难度

合理地综合使用信息元素：图>曲线>表>正文>公式

尽量提高读者阅读时的愉悦感

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

尽量降低读者的理解难度

合理地综合使用信息元素：图>曲线>表>正文>公式

尽量提高读者阅读时的愉悦感

思想新颖、组织合理、逻辑严密
论证充分、文笔优美、排版美观

摘要的写作技巧

摘要

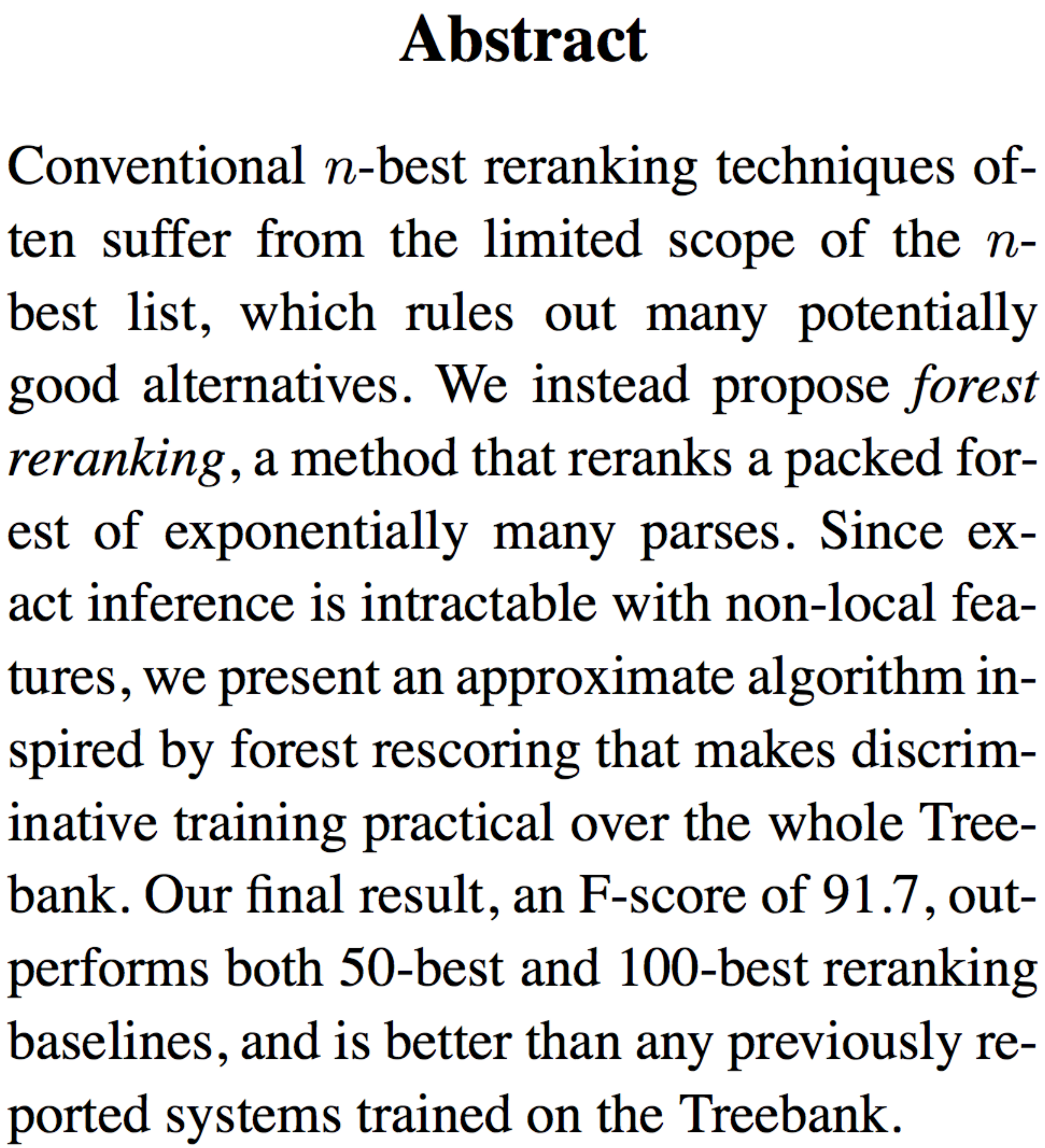
- 几句话概括你的工作
- 误区
  - 力图把所有细节都说清楚
  - 用很专业的术语来描述
  - 出现数学符号

摘要

- 几句话概括你的工作
- 误区
  - 力图把所有细节都说清楚
  - 用很专业的术语来描述
  - 出现数学符号

用语要简单，让外行能看懂

例子

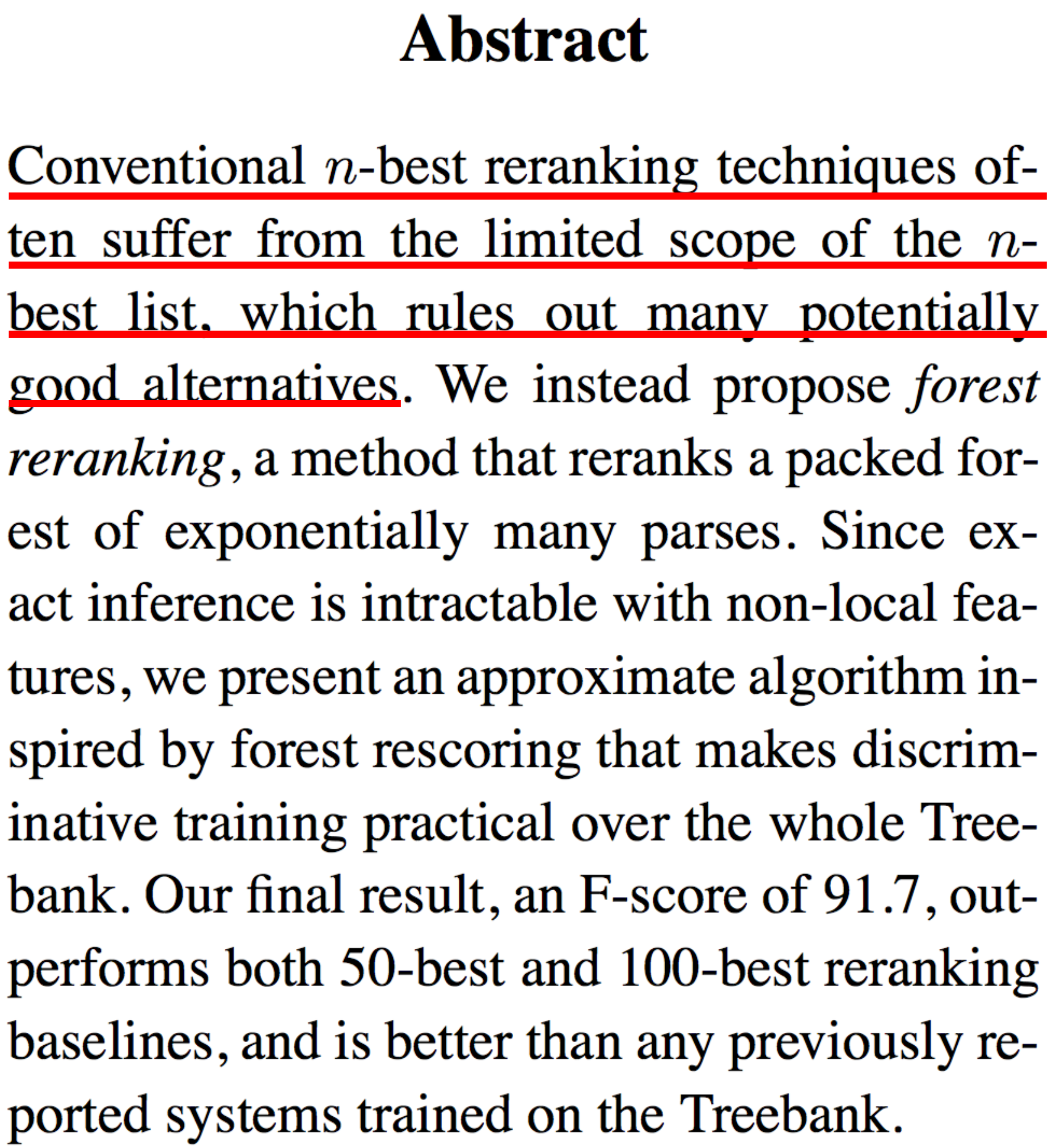


Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子



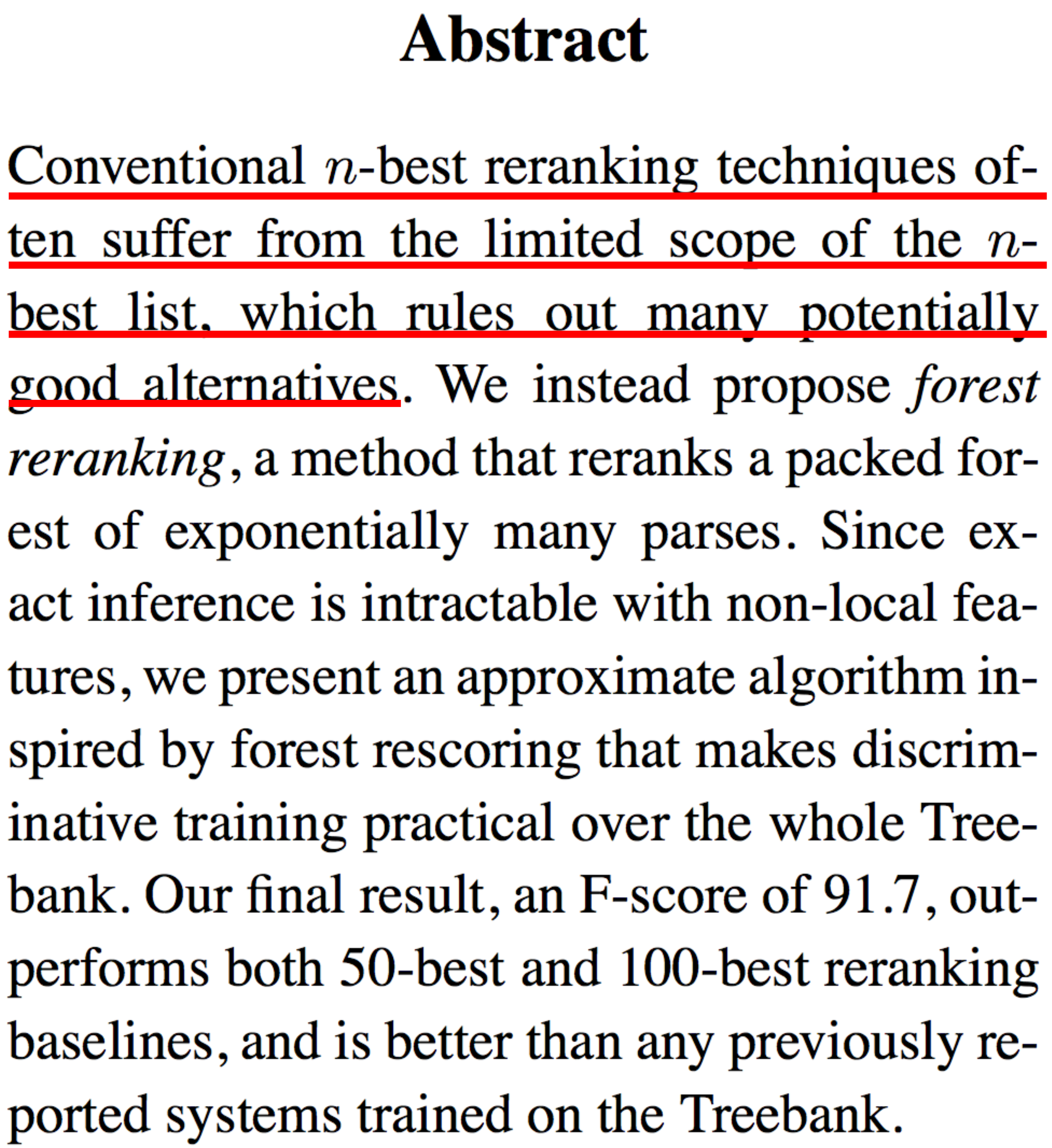
Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子

问题是什么



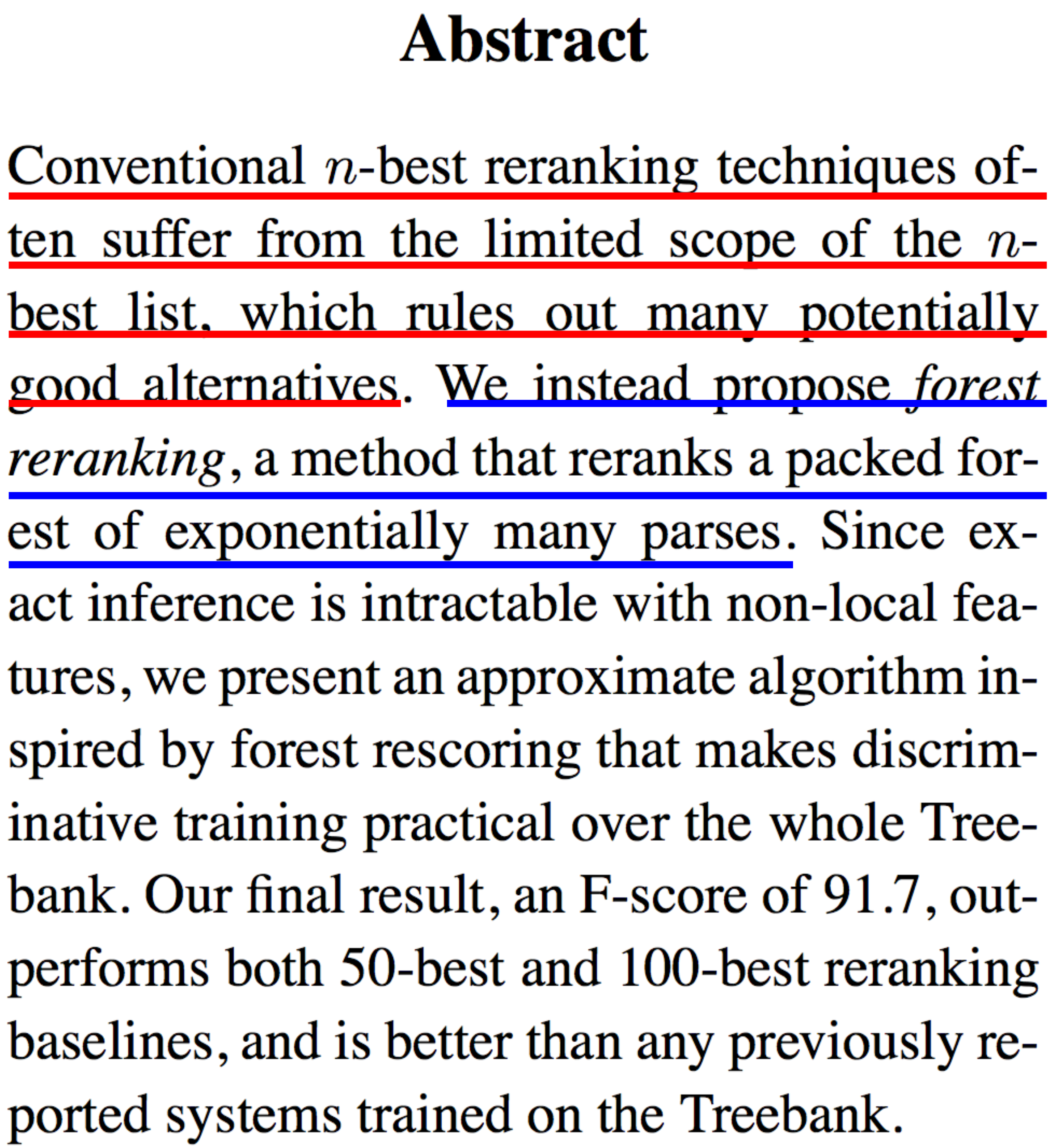
Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子

问题是什么



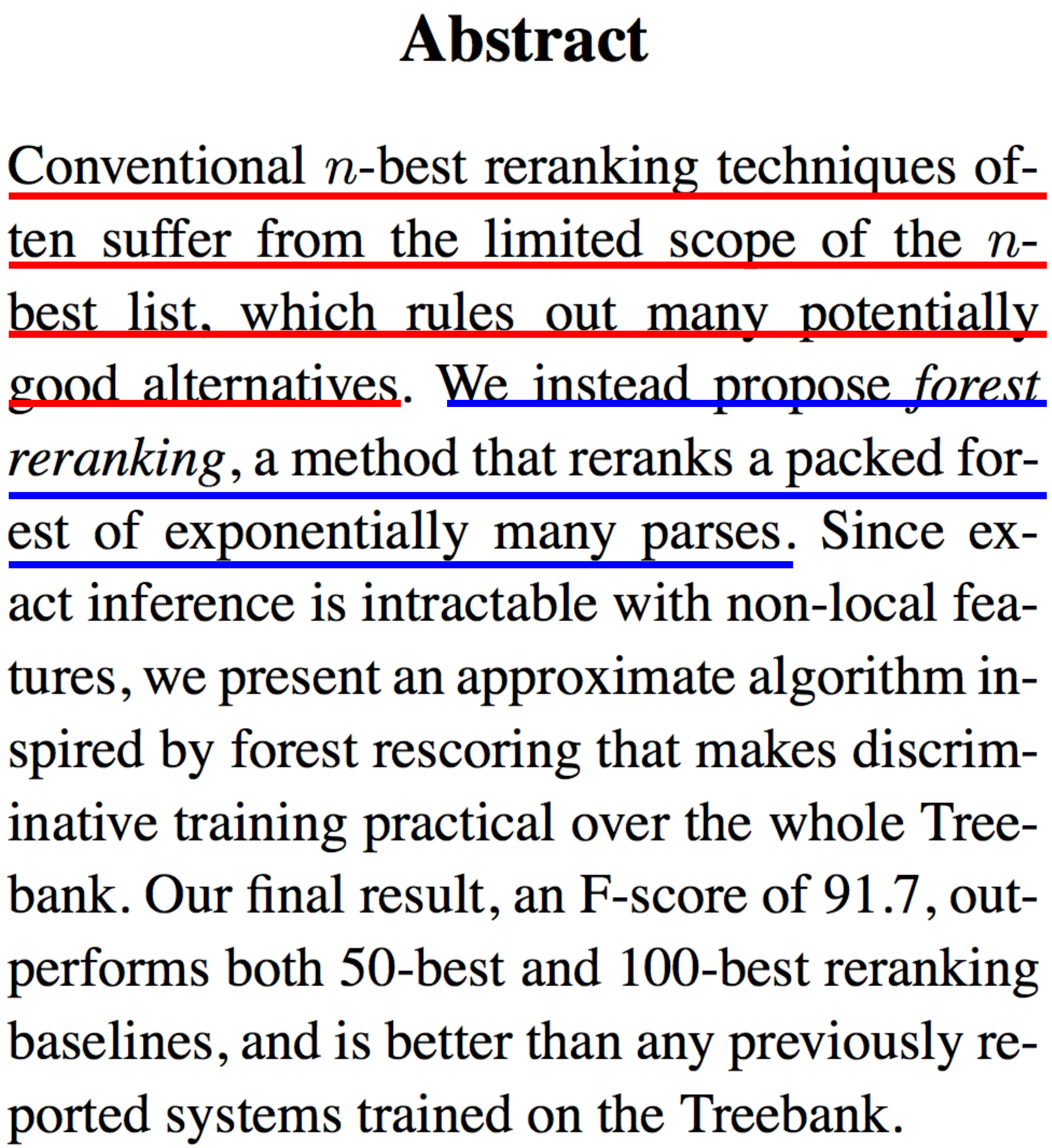
Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子

问题是什么



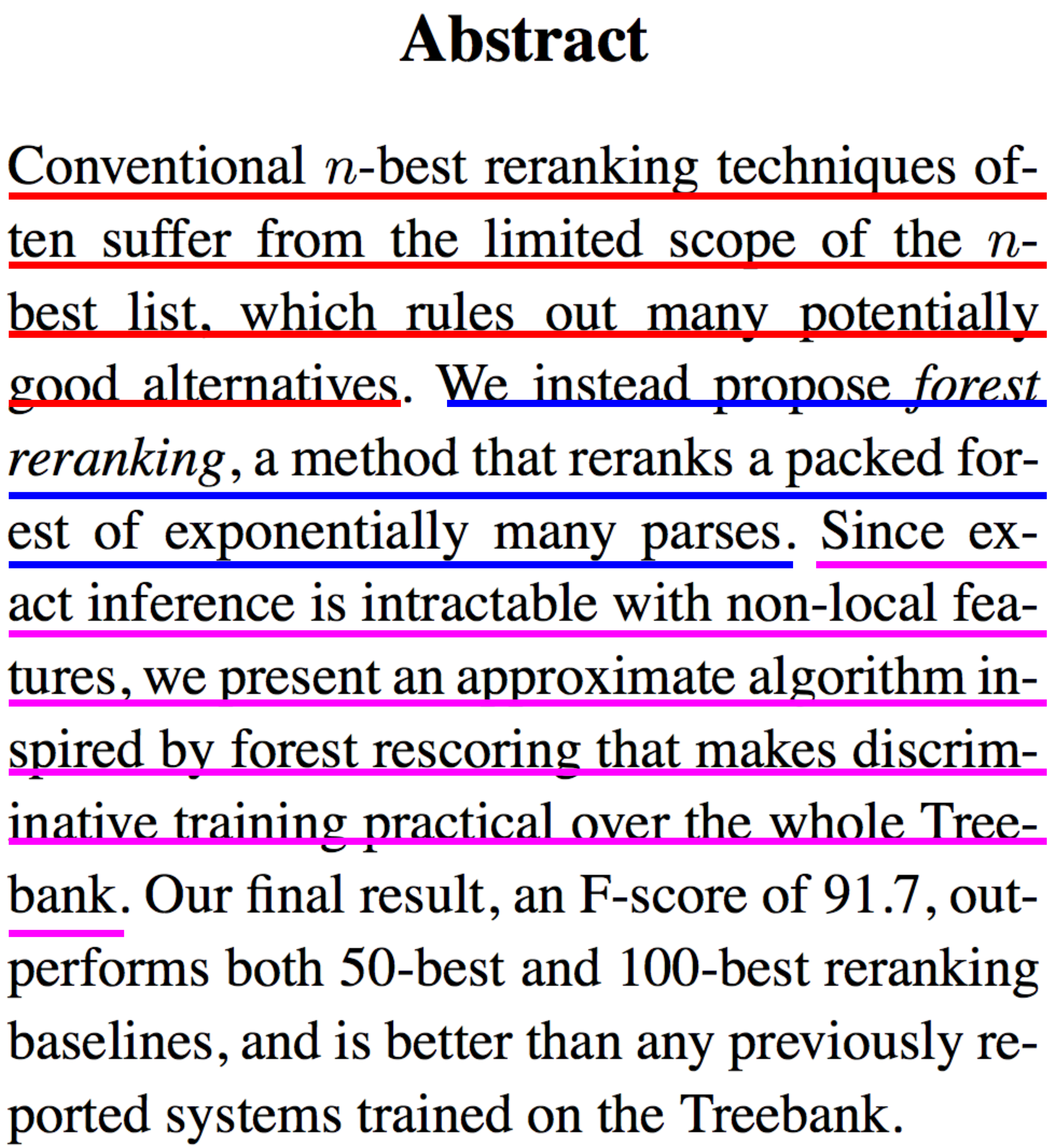
Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

我们做了什么

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子



Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

问题是什么

我们做了什么

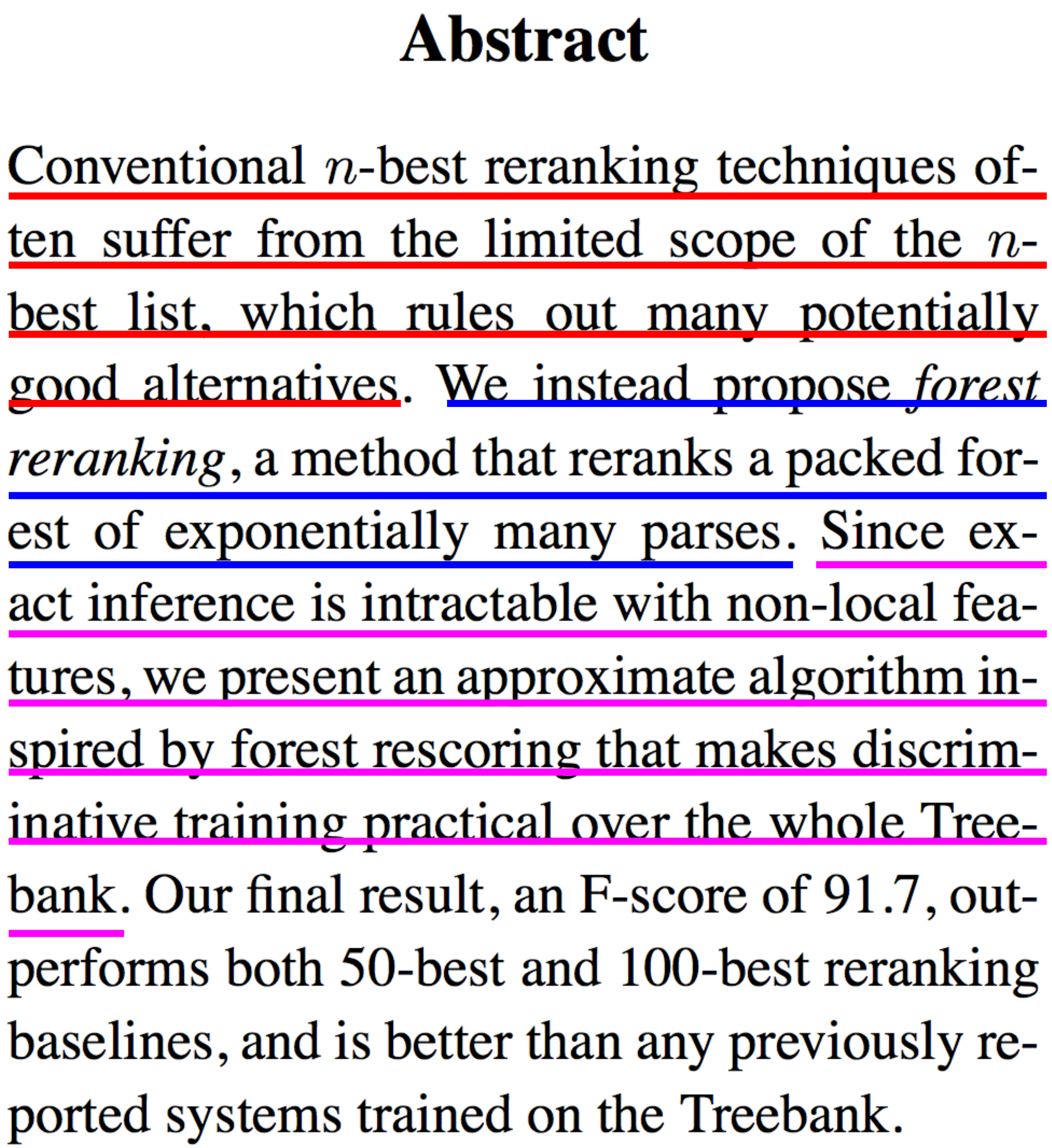
Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子

问题是什么

我们做了什么

我们大概怎么做的

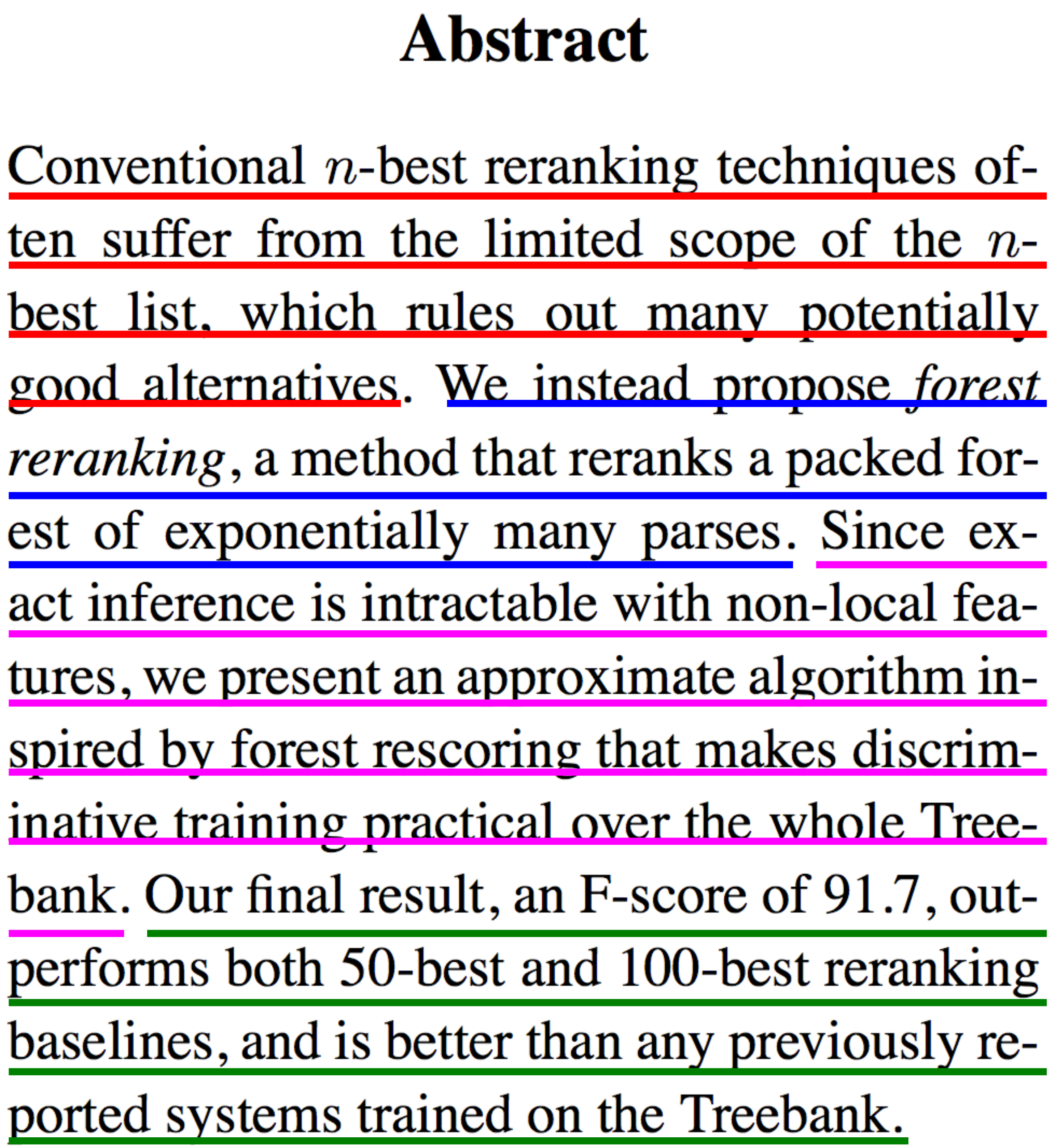


Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子



Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

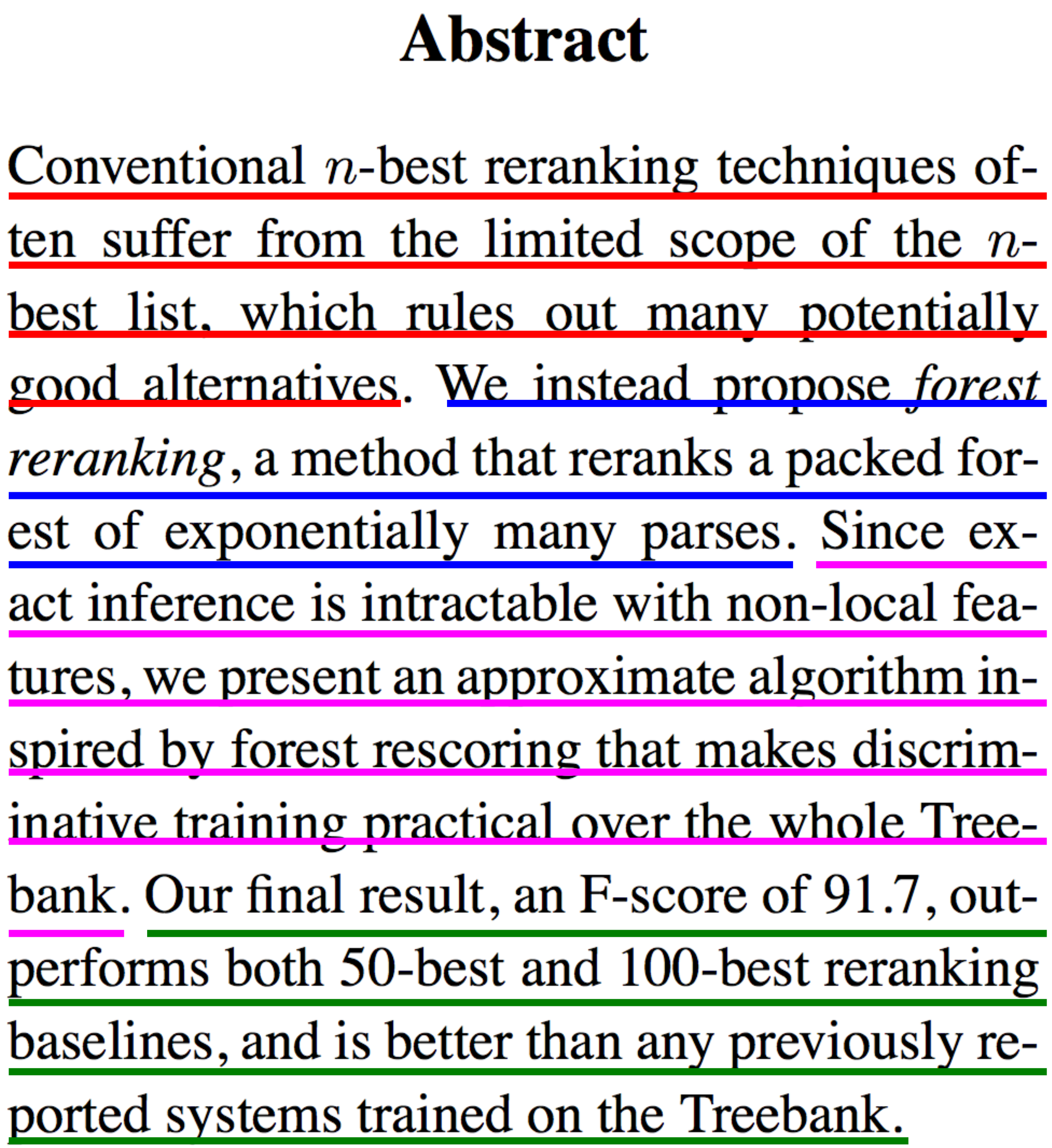
问题是什么

我们做了什么

我们大概怎么做的

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

例子



Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

问题是什么

我们做了什么

我们大概怎么做的

我们做得挺不错!

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

介绍的写作技巧

介绍的写法

- 比题目和摘要更进一步，用几段话说清你的工作

- 要点是充分论证你所做工作的必要性和重要性，要让审稿人认同并迫不及待想往下看。

- 行文逻辑严密，论证充分

逻辑

- 常见的逻辑
  - 说明问题是什么
  - 简单罗列前人工作
  - 描述我们的工作

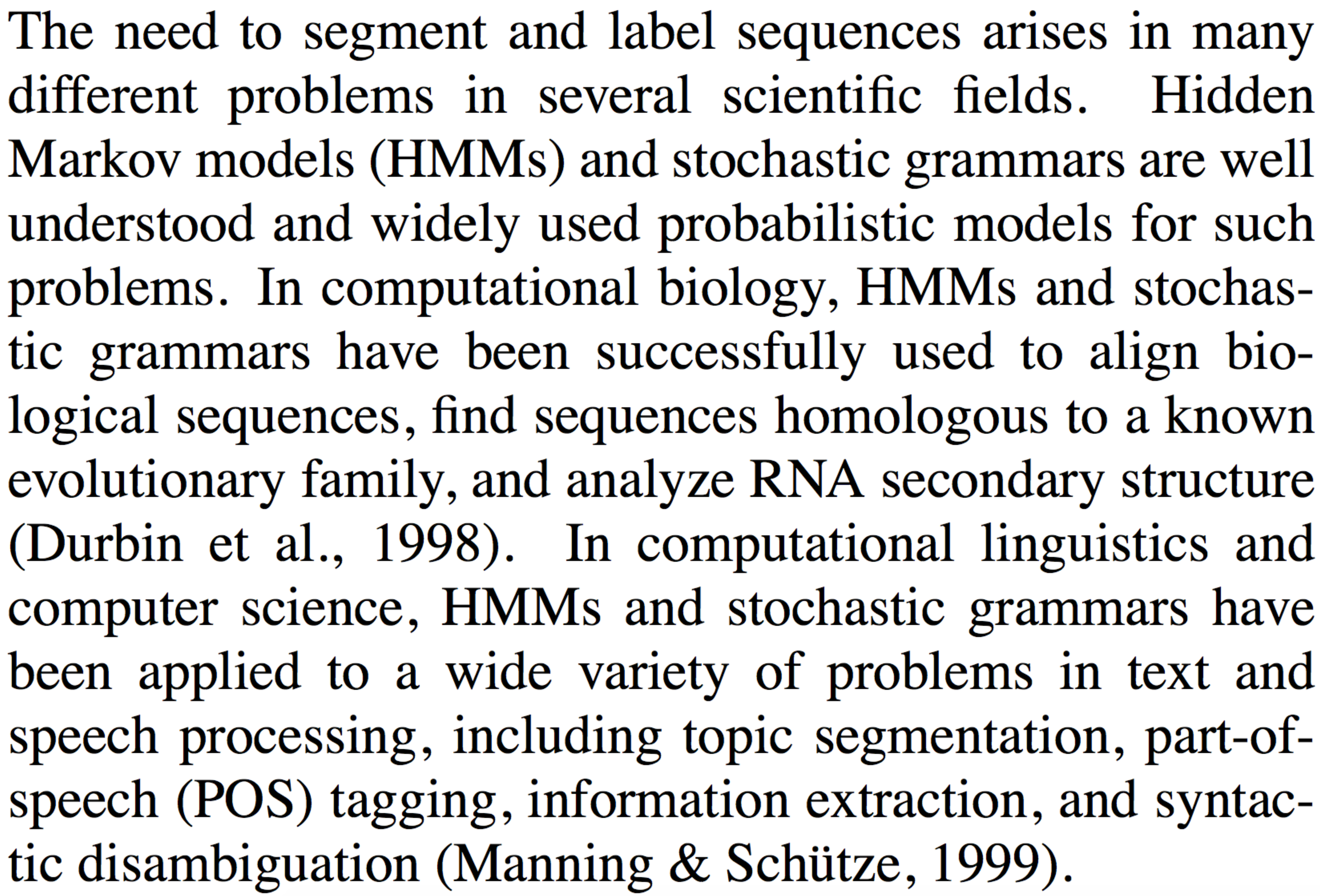
逻辑

- 常见的逻辑
  - 说明问题是什么
  - 简单罗列前人工作
  - 描述我们的工作
- 更好的逻辑
  - 说明问题是什么
  - 目前最好的工作面临什么挑战
  - 我们的方法能缓解上述挑战

段落的写法

- 每个段落有个论断性的中心句
- 其余部分都是支撑句，围绕中心句展开论证
  - 前人工作
  - 具体数据
- 支撑句之间可分类组织
- 段尾可以加上衔接句

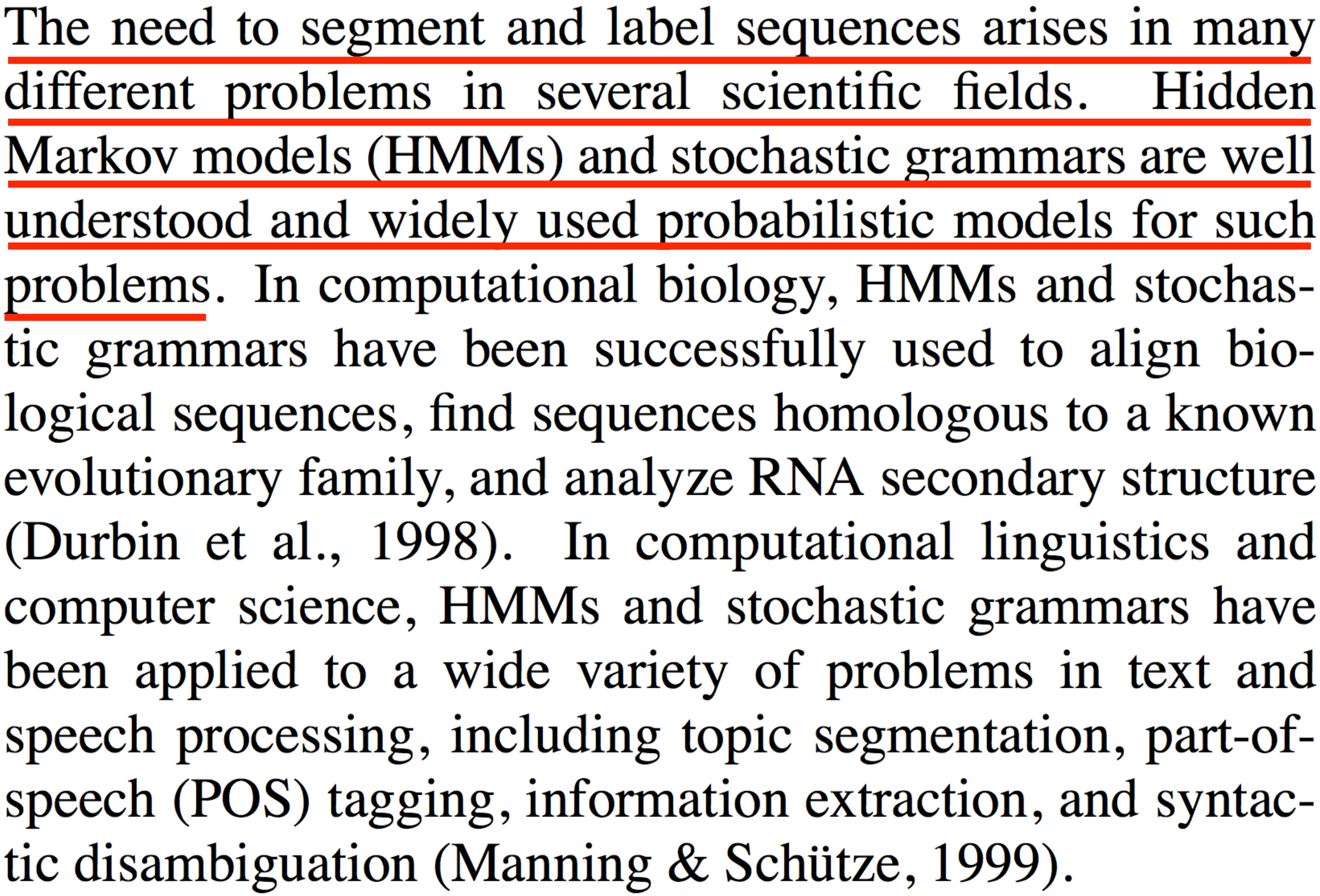
中心句与支撑句



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. In computational biology, HMMs and stochastic grammars have been successfully used to align biological sequences, find sequences homologous to a known evolutionary family, and analyze RNA secondary structure (Durbin et al., 1998). In computational linguistics and computer science, HMMs and stochastic grammars have been applied to a wide variety of problems in text and speech processing, including topic segmentation, part-of-speech (POS) tagging, information extraction, and syntactic disambiguation (Manning & Schütze, 1999).

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

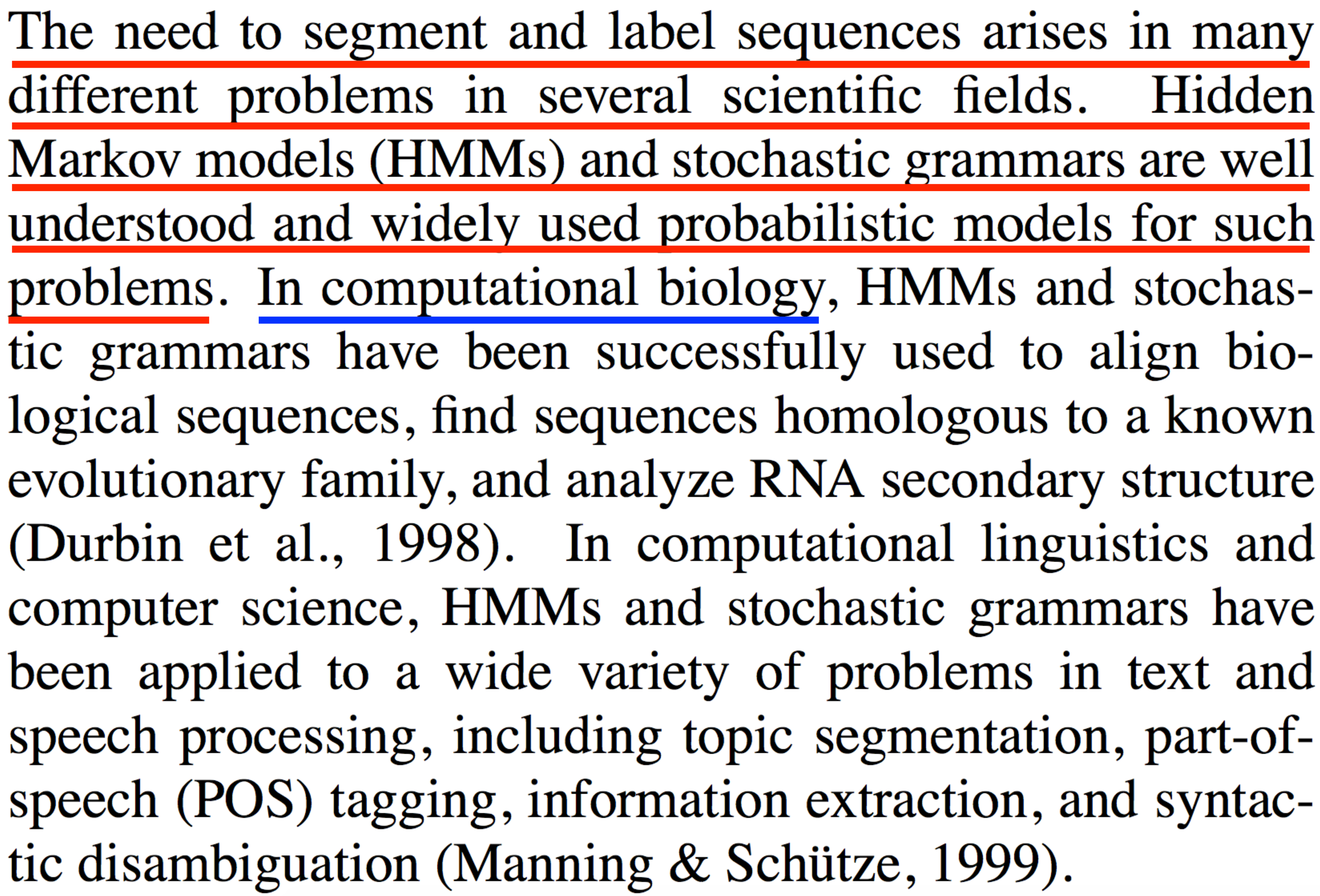
中心句与支撑句



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. In computational biology, HMMs and stochastic grammars have been successfully used to align biological sequences, find sequences homologous to a known evolutionary family, and analyze RNA secondary structure (Durbin et al., 1998). In computational linguistics and computer science, HMMs and stochastic grammars have been applied to a wide variety of problems in text and speech processing, including topic segmentation, part-of-speech (POS) tagging, information extraction, and syntactic disambiguation (Manning & Schütze, 1999).

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

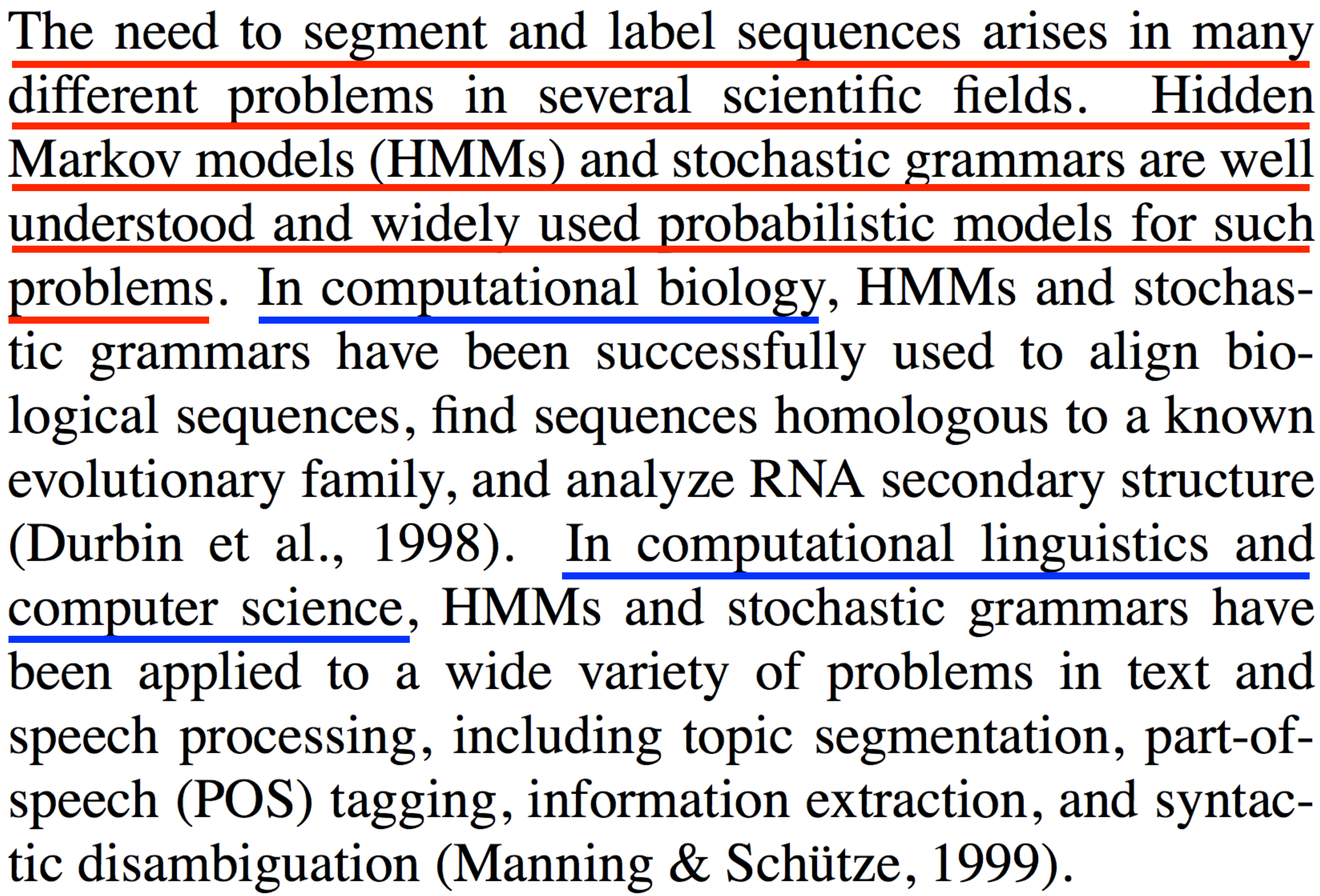
中心句与支撑句



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. In computational biology, HMMs and stochastic grammars have been successfully used to align biological sequences, find sequences homologous to a known evolutionary family, and analyze RNA secondary structure (Durbin et al., 1998). In computational linguistics and computer science, HMMs and stochastic grammars have been applied to a wide variety of problems in text and speech processing, including topic segmentation, part-of-speech (POS) tagging, information extraction, and syntactic disambiguation (Manning & Schütze, 1999).

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

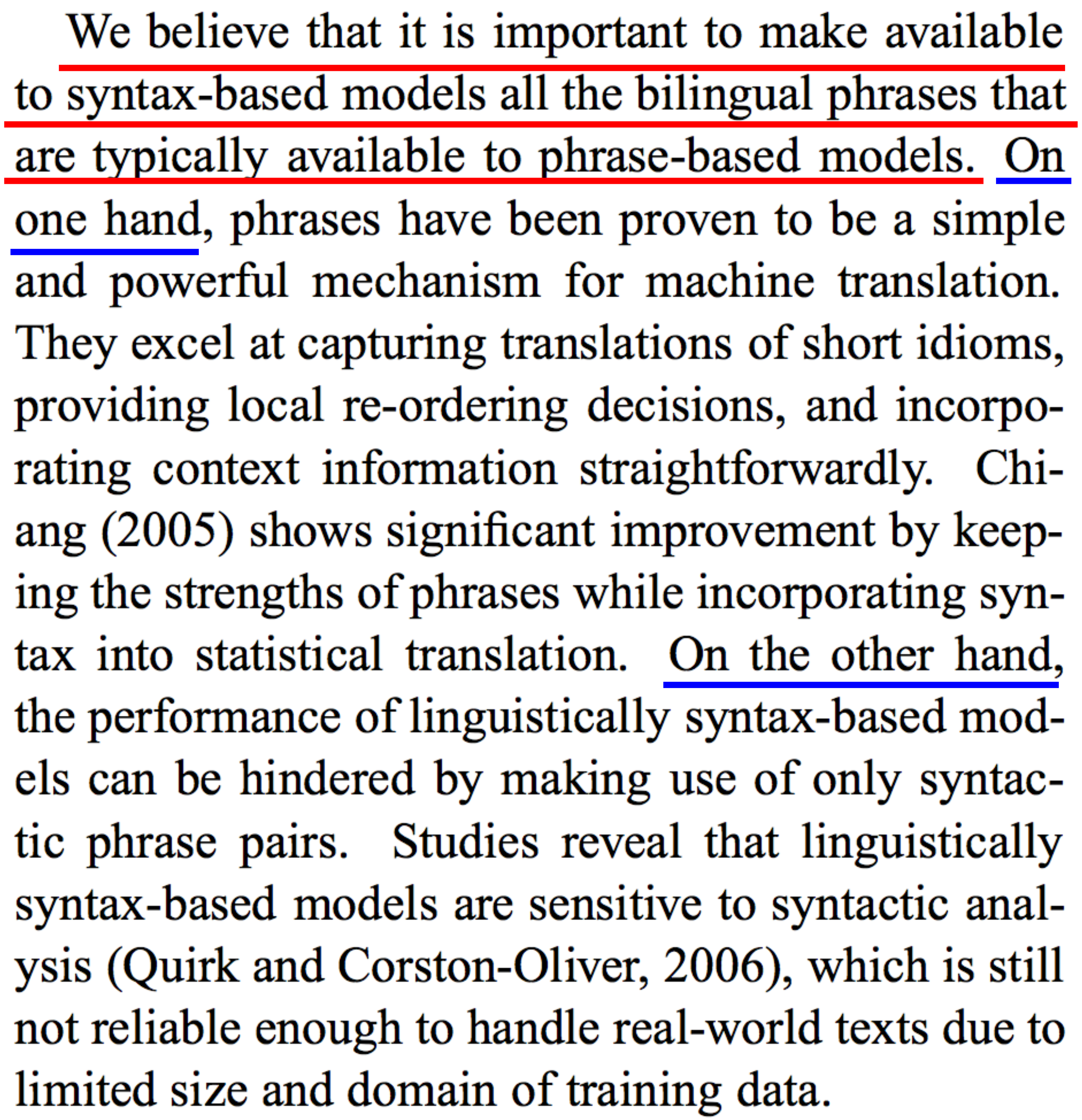
中心句与支撑句



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. In computational biology, HMMs and stochastic grammars have been successfully used to align biological sequences, find sequences homologous to a known evolutionary family, and analyze RNA secondary structure (Durbin et al., 1998). In computational linguistics and computer science, HMMs and stochastic grammars have been applied to a wide variety of problems in text and speech processing, including topic segmentation, part-of-speech (POS) tagging, information extraction, and syntactic disambiguation (Manning & Schütze, 1999).

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

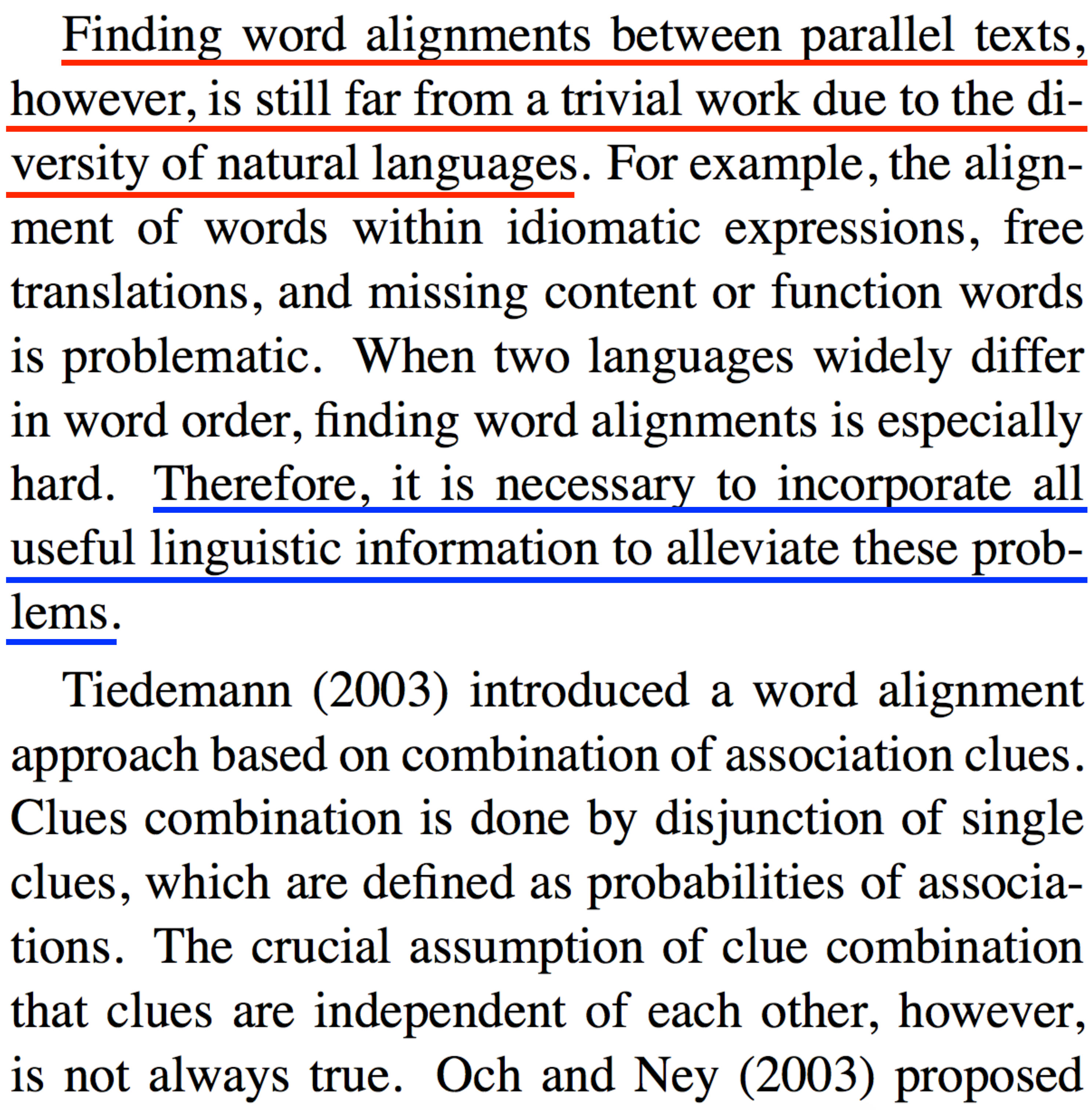
中心句与支撑句



We believe that it is important to make available to syntax-based models all the bilingual phrases that are typically available to phrase-based models. On one hand, phrases have been proven to be a simple and powerful mechanism for machine translation. They excel at capturing translations of short idioms, providing local re-ordering decisions, and incorporating context information straightforwardly. Chiang (2005) shows significant improvement by keeping the strengths of phrases while incorporating syntax into statistical translation. On the other hand, the performance of linguistically syntax-based models can be hindered by making use of only syntactic phrase pairs. Studies reveal that linguistically syntax-based models are sensitive to syntactic analysis (Quirk and Corston-Oliver, 2006), which is still not reliable enough to handle real-world texts due to limited size and domain of training data.

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

衔接句

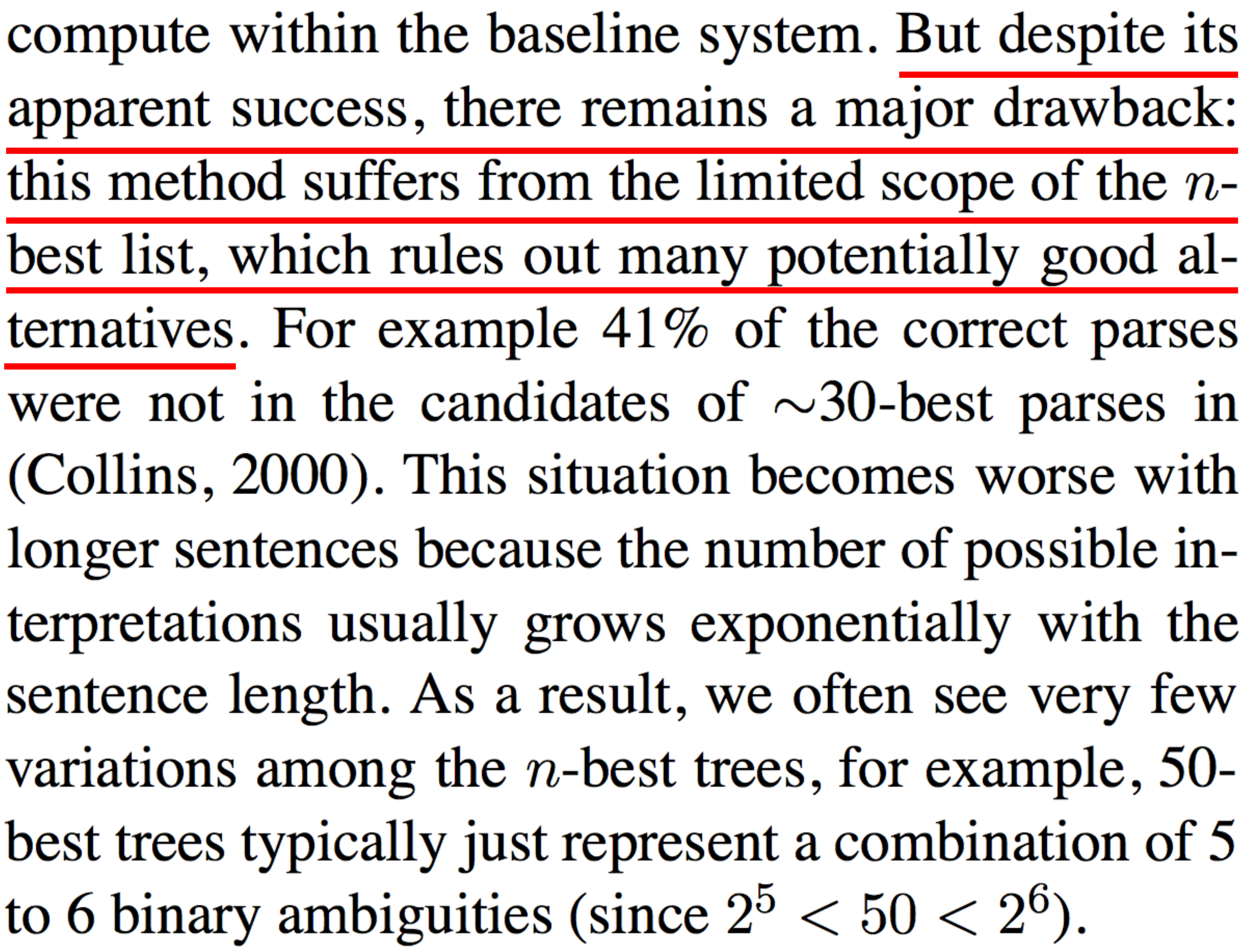


Finding word alignments between parallel texts, however, is still far from a trivial work due to the diversity of natural languages. For example, the alignment of words within idiomatic expressions, free translations, and missing content or function words is problematic. When two languages widely differ in word order, finding word alignments is especially hard. Therefore, it is necessary to incorporate all useful linguistic information to alleviate these problems.

Tiedemann (2003) introduced a word alignment approach based on combination of association clues. Clues combination is done by disjunction of single clues, which are defined as probabilities of associations. The crucial assumption of clue combination that clues are independent of each other, however, is not always true. Och and Ney (2003) proposed

Yang Liu, Qun Liu, and Shogun Lin. **Log-Linear Models for Word Alignment**. In *ACL 2005*.

支撑句要论证严密



compute within the baseline system. But despite its apparent success, there remains a major drawback: this method suffers from the limited scope of the n-best list, which rules out many potentially good alternatives. For example 41% of the correct parses were not in the candidates of $\sim$30-best parses in (Collins, 2000). This situation becomes worse with longer sentences because the number of possible interpretations usually grows exponentially with the sentence length. As a result, we often see very few variations among the n-best trees, for example, 50-best trees typically just represent a combination of 5 to 6 binary ambiguities (since $2^5 < 50 < 2^6$).

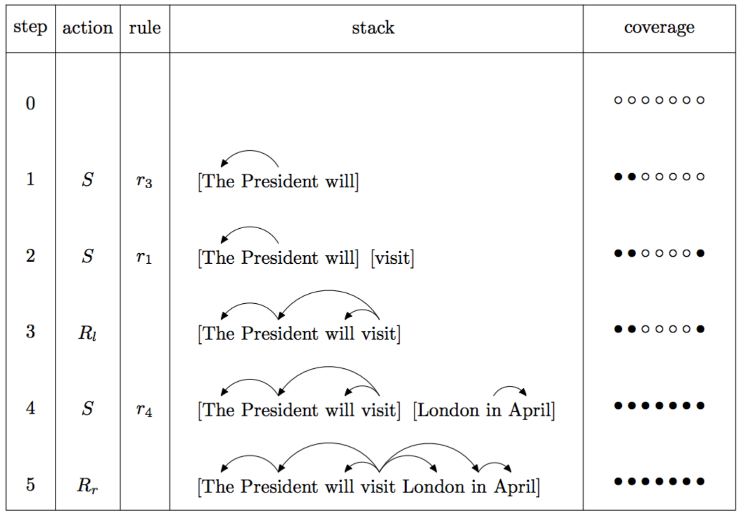
Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

新技巧

- 在首页放置一个图或者表，让读者一目了然你所做的工作；
- 不要去写“This paper is organized as follows. Section 2 …”，而是直接列出自己的贡献。

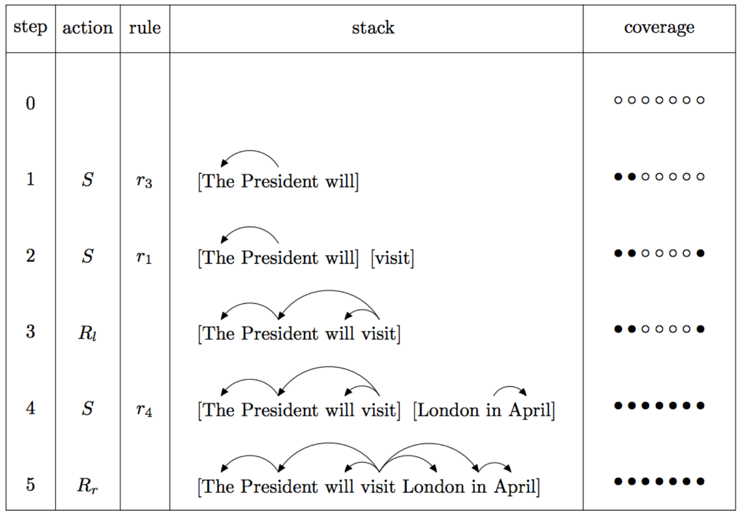
信息元素的易理解度

信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

信息元素的易理解度

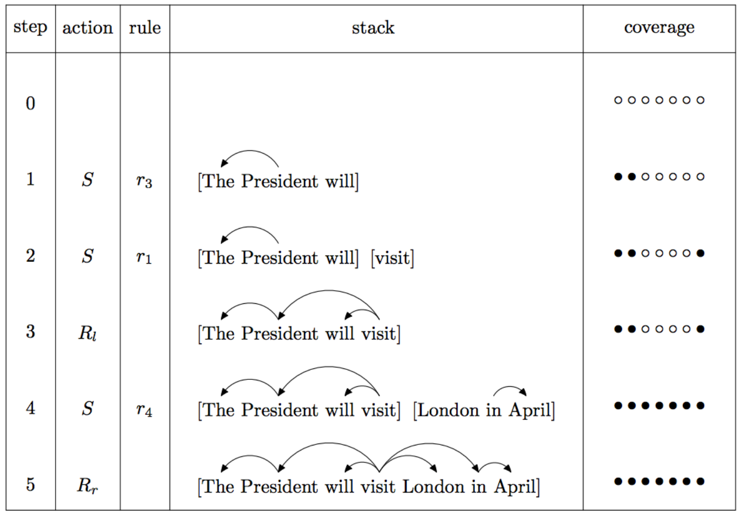


step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

*

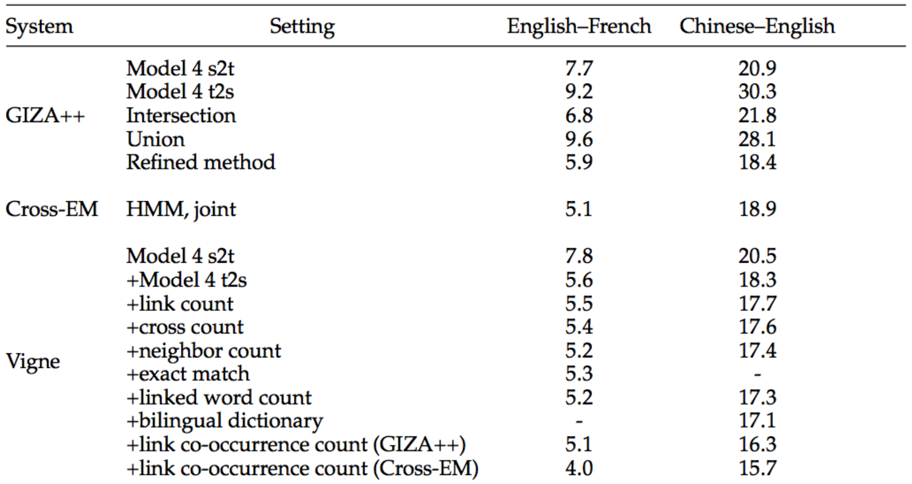
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

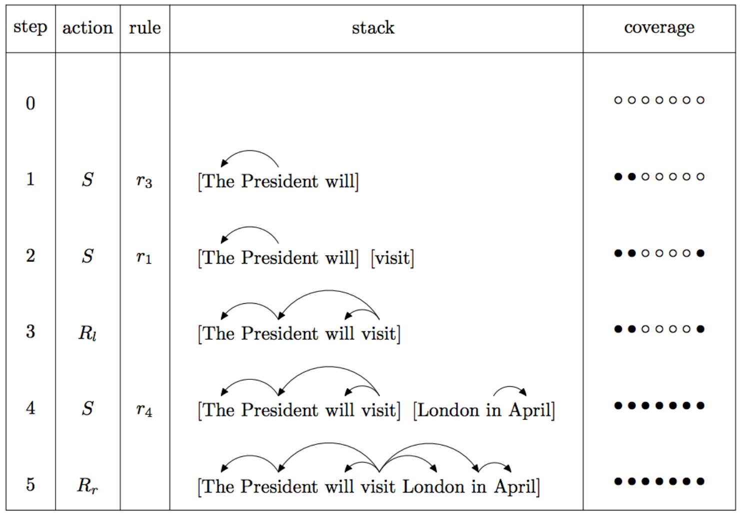
图

*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

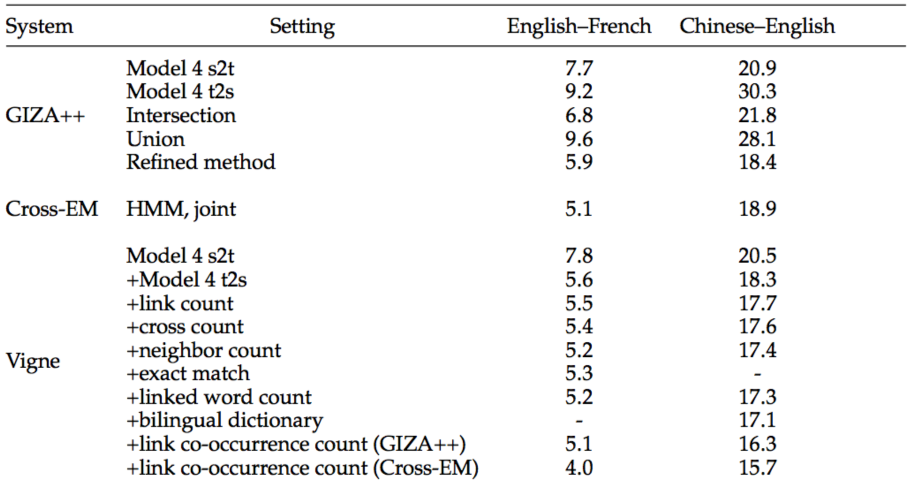
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

*

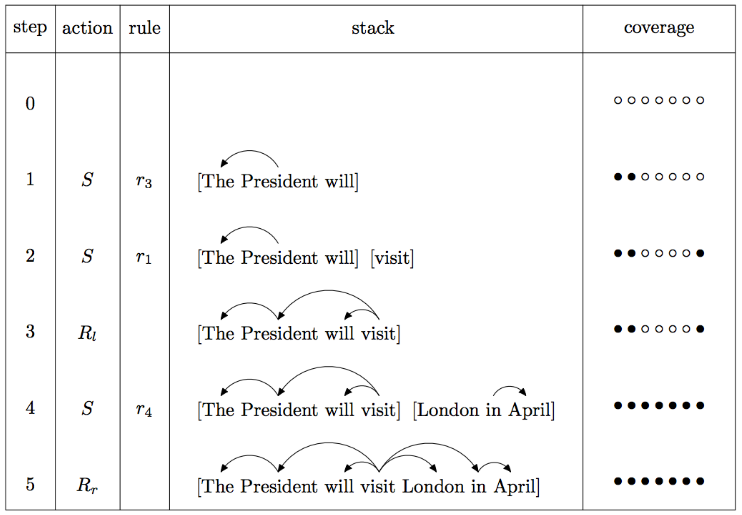


System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

**

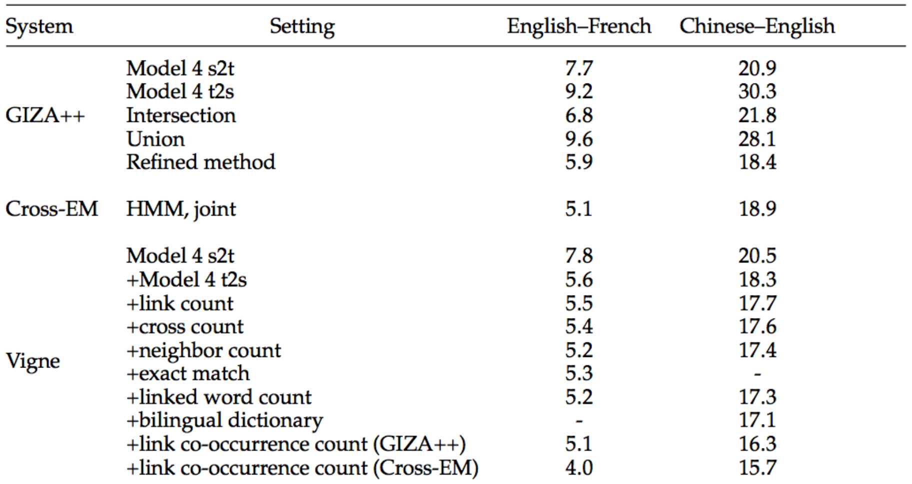
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

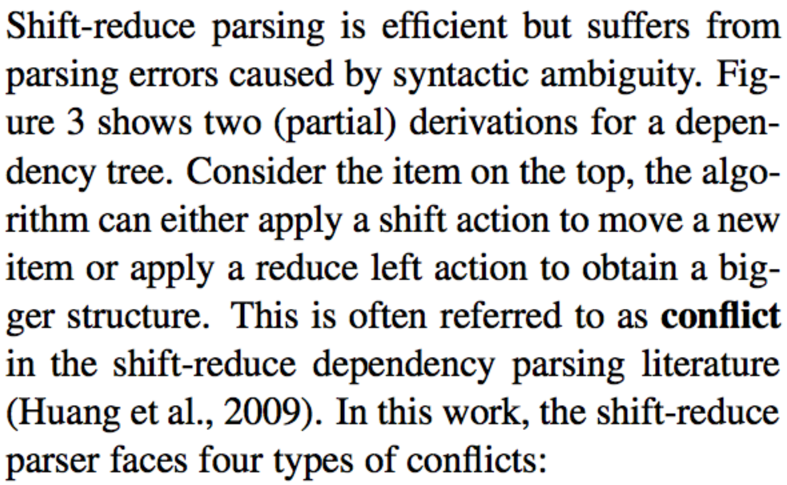
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

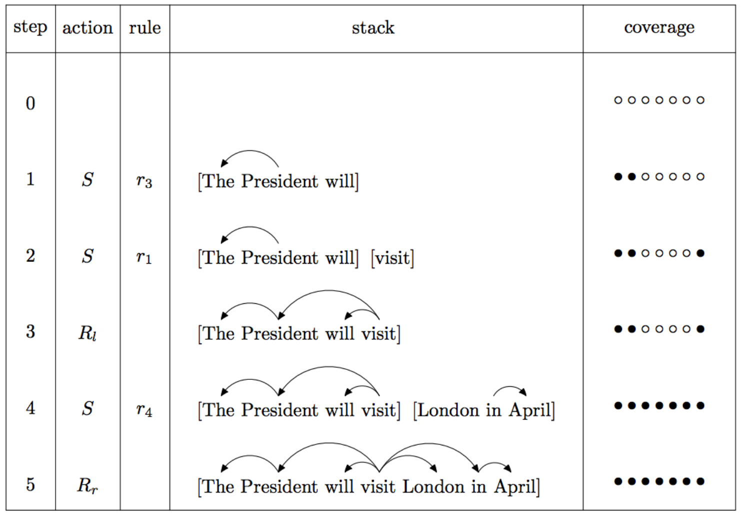
表格

**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

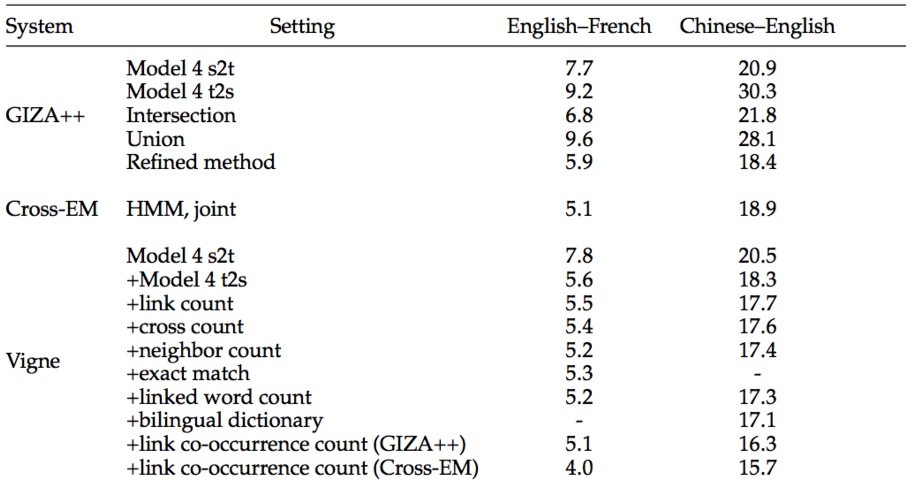
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

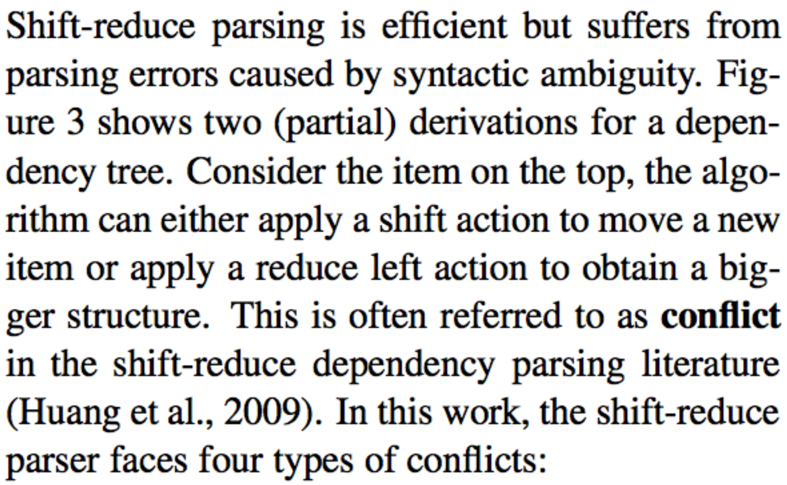
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

**

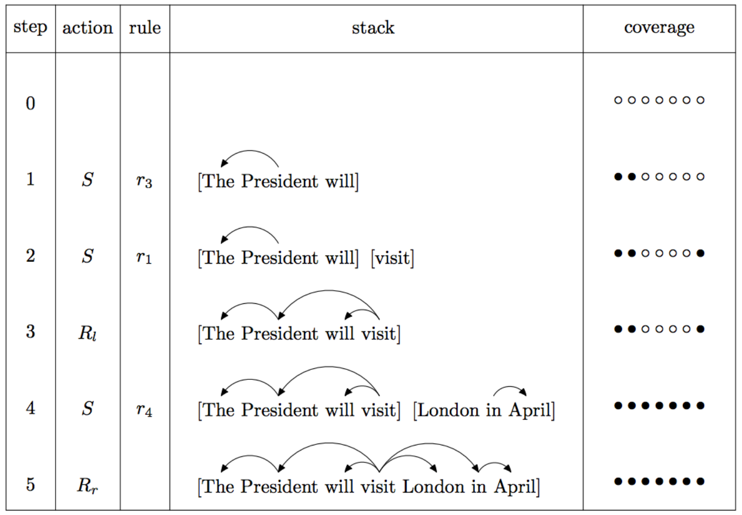


Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

***

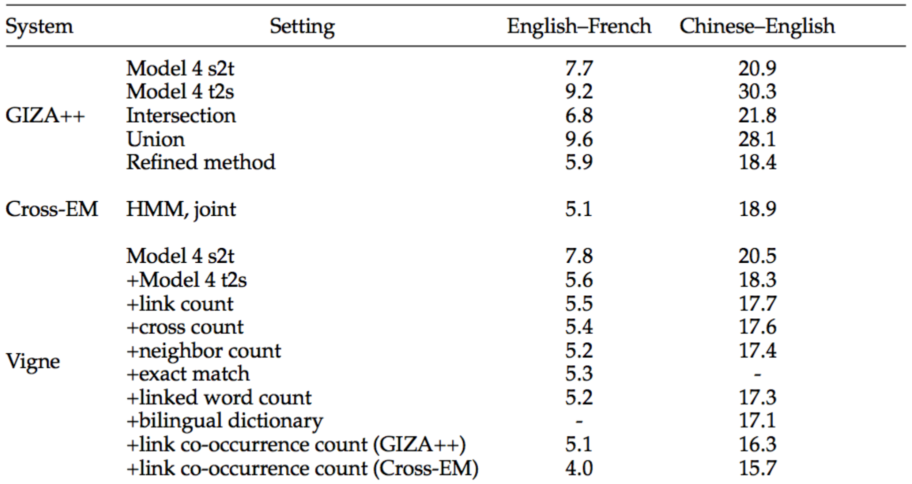
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

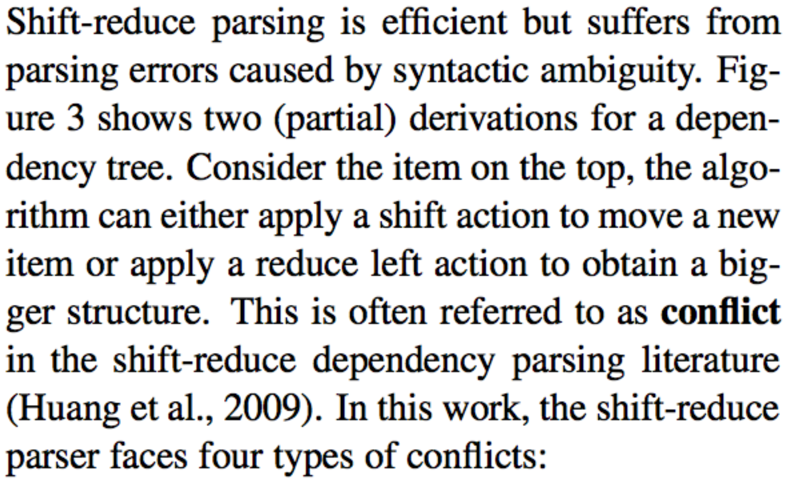
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

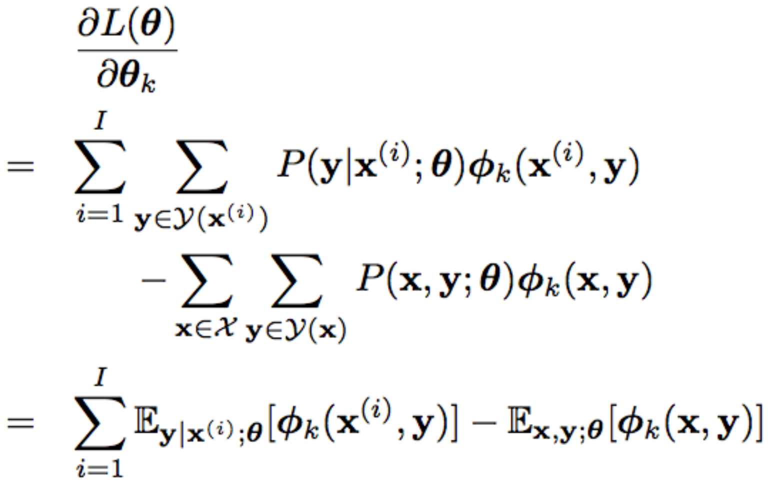
**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

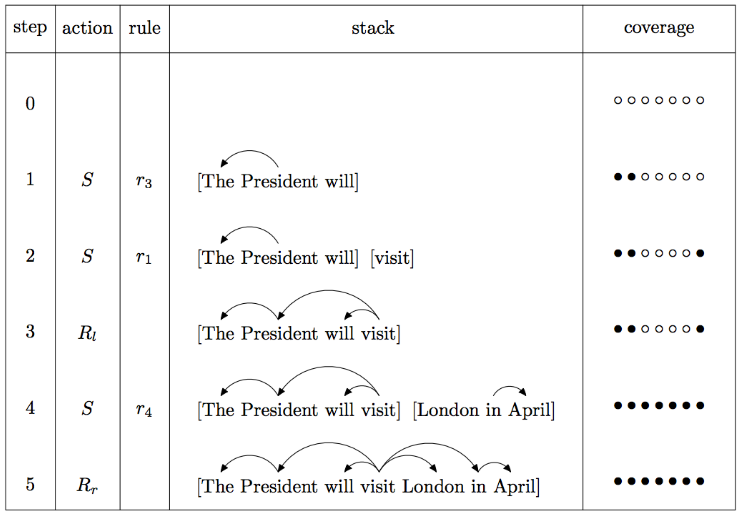
正文

***



$$
\begin{aligned}
&\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \\
&= \sum_{i=1}^{I} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)}; \boldsymbol{\theta}) \phi_k(\mathbf{x}^{(i)}, \mathbf{y}) \\
&\qquad - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} P(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \phi_k(\mathbf{x}, \mathbf{y}) \\
&= \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)}, \mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x}, \mathbf{y})]
\end{aligned}
$$

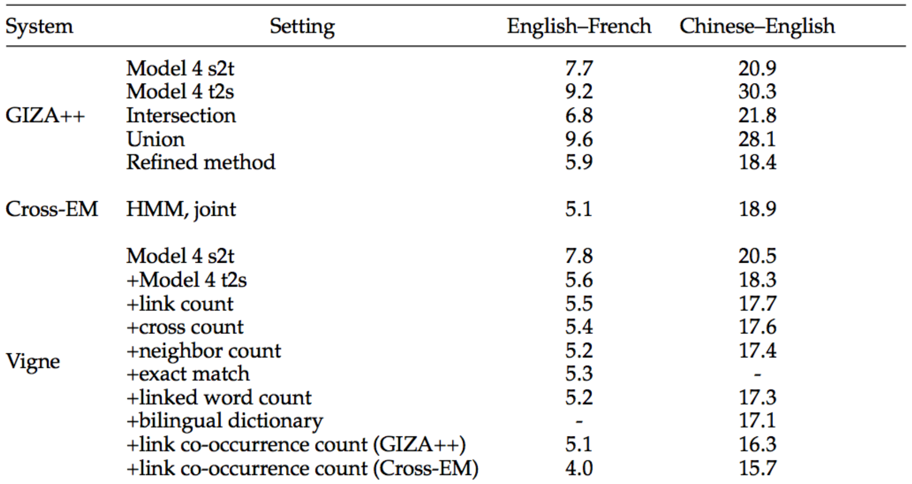
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

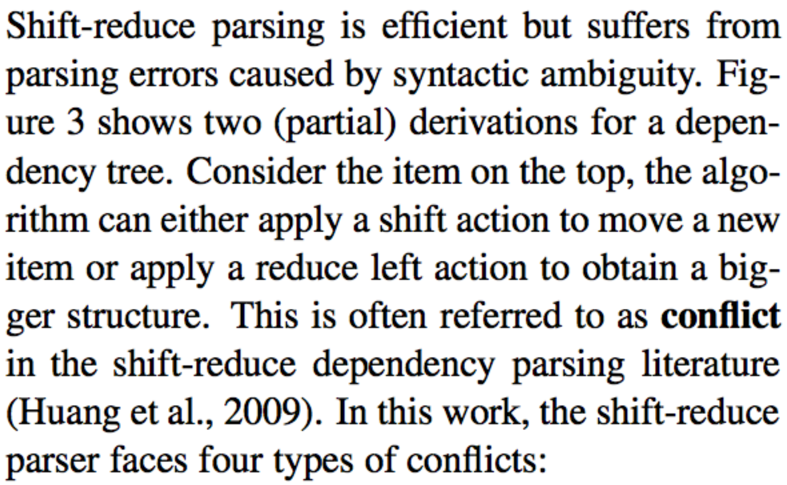
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

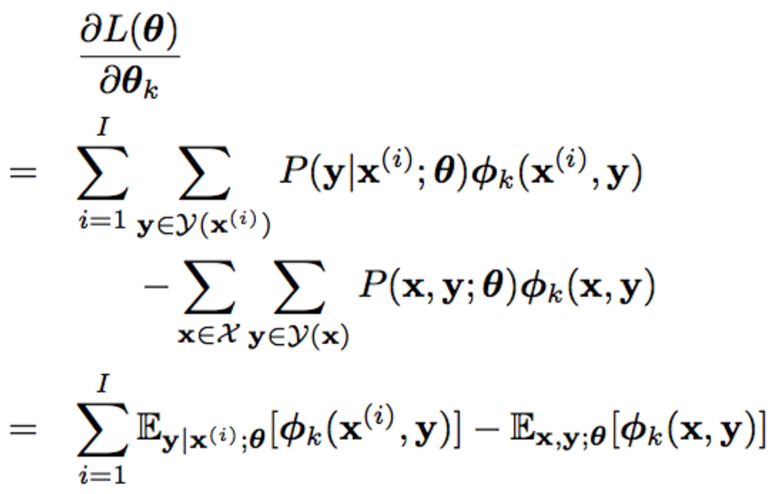
**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

***

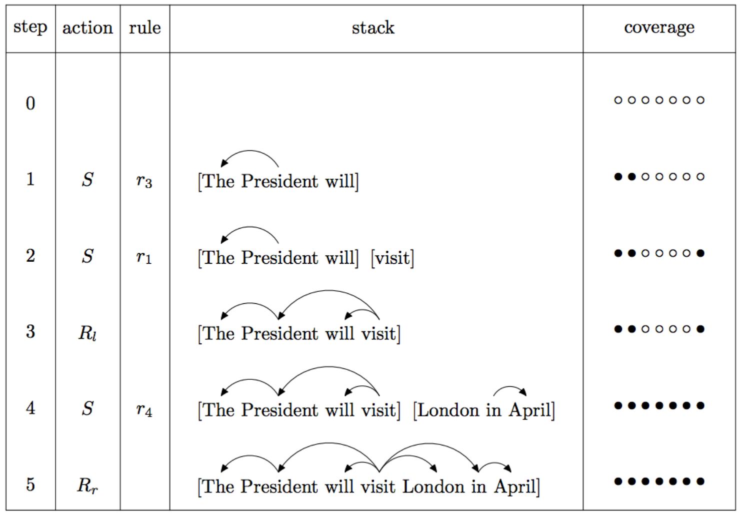


$$\begin{aligned}
&\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \\
= &\sum_{i=1}^{I} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta})\phi_k(\mathbf{x}^{(i)},\mathbf{y}) \\
&\quad - \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\phi_k(\mathbf{x},\mathbf{y}) \\
= &\sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x},\mathbf{y})]
\end{aligned}$$

公式

****

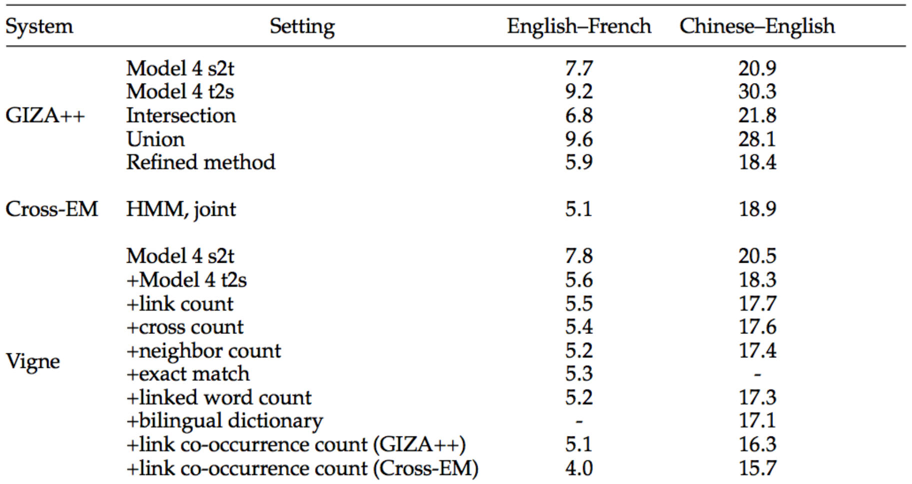
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

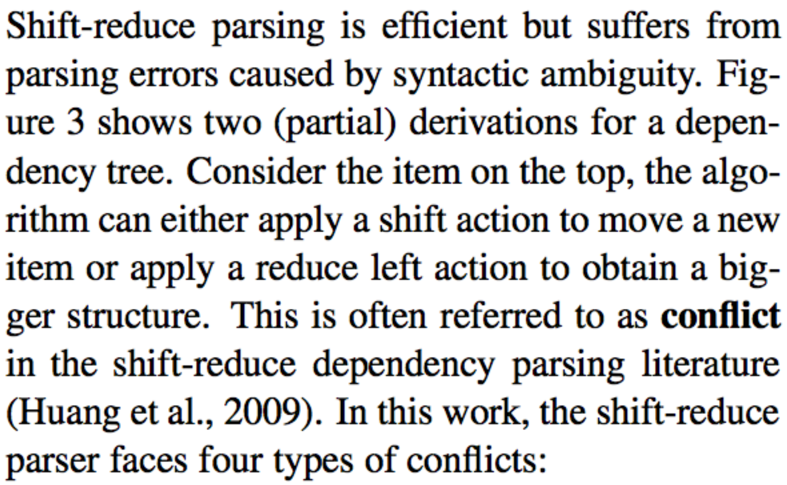
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

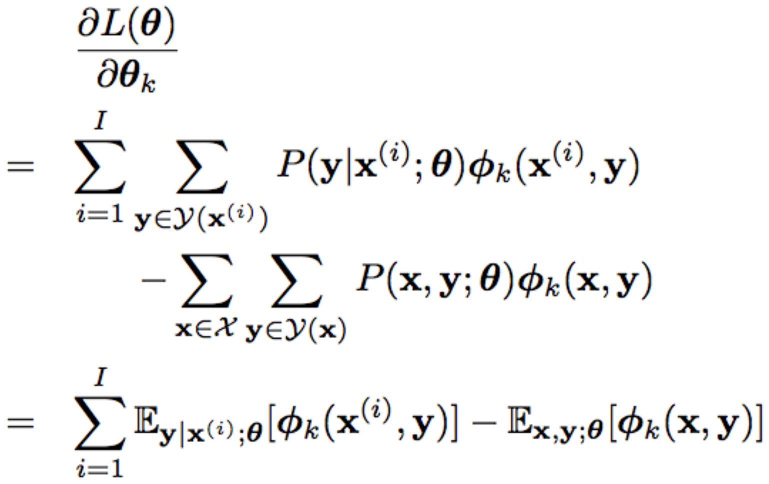
**



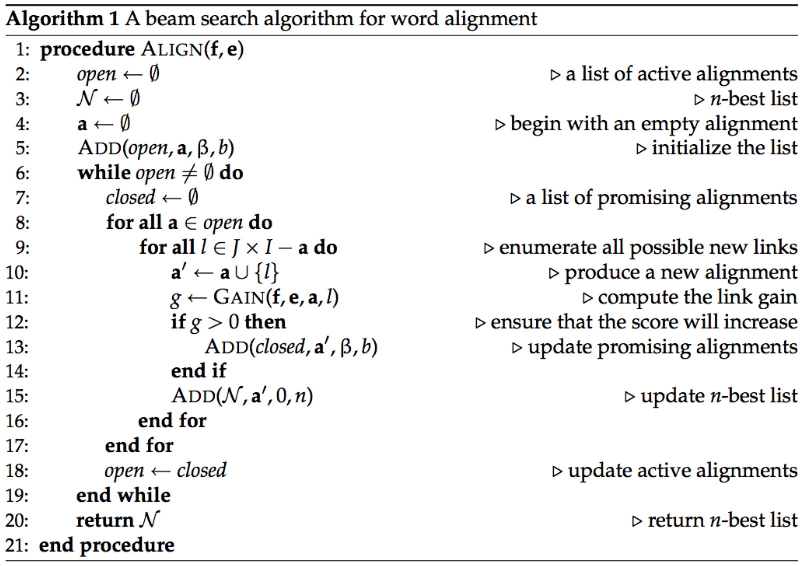
Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

***



$$\begin{aligned}
&\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_k} \\
&= \sum_{i=1}^{I} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta})\phi_k(\mathbf{x}^{(i)},\mathbf{y}) \\
&\qquad - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\phi_k(\mathbf{x},\mathbf{y}) \\
&= \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x},\mathbf{y})]
\end{aligned}$$



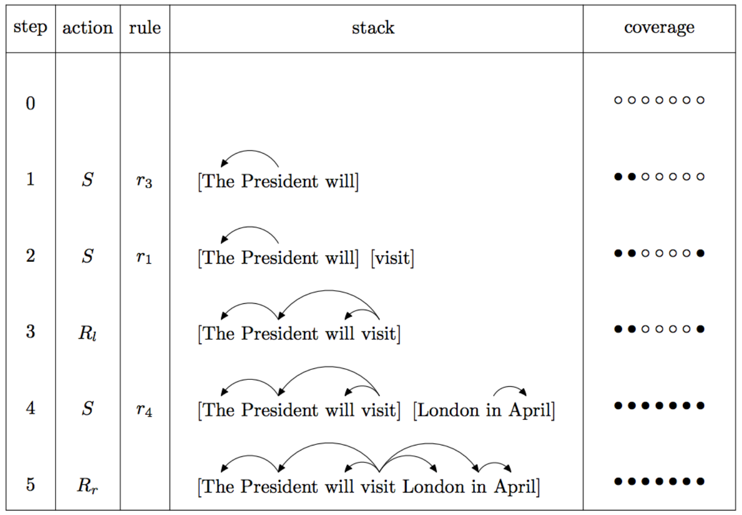
Algorithm 1 A beam search algorithm for word alignment

```
procedure ALIGN($\mathbf{f}, \mathbf{e}$)
    $open \leftarrow \emptyset$                                  ▷ a list of active alignments
    $\mathcal{N} \leftarrow \emptyset$                                  ▷ $n$-best list
    $\mathbf{a} \leftarrow \emptyset$                                  ▷ begin with an empty alignment
    ADD($open, \mathbf{a}, \beta, b$)                        ▷ initialize the list
    while $open \neq \emptyset$ do
        $closed \leftarrow \emptyset$                            ▷ a list of promising alignments
        for all $\mathbf{a} \in open$ do
            for all $l \in J \times I - \mathbf{a}$ do           ▷ enumerate all possible new links
                $\mathbf{a}' \leftarrow \mathbf{a} \cup \{l\}$                  ▷ produce a new alignment
                $g \leftarrow$ GAIN($\mathbf{f}, \mathbf{e}, \mathbf{a}, l$)              ▷ compute the link gain
                if $g > 0$ then                         ▷ ensure that the score will increase
                    ADD($closed, \mathbf{a}', \beta, b$)         ▷ update promising alignments
                end if
                ADD($\mathcal{N}, \mathbf{a}', 0, n$)                 ▷ update $n$-best list
            end for
        end for
        $open \leftarrow closed$                         ▷ update active alignments
    end while
    return $\mathcal{N}$                                  ▷ return $n$-best list
end procedure
```

公式

****

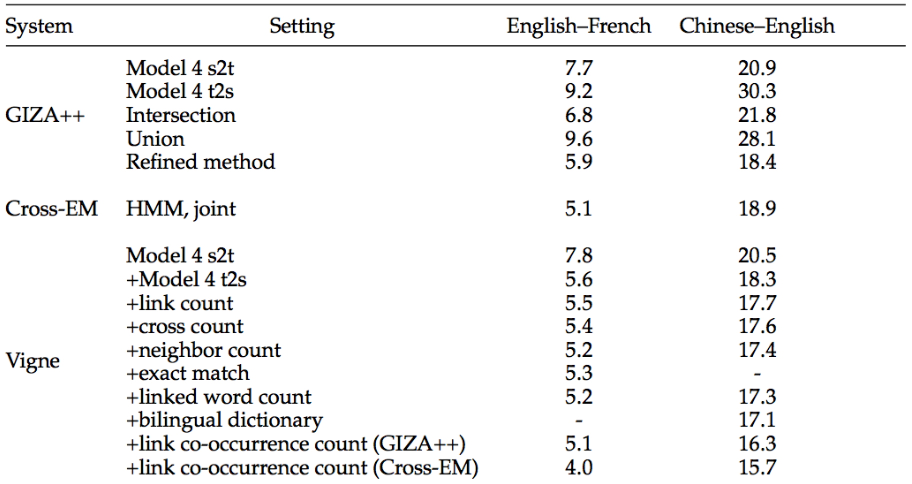
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

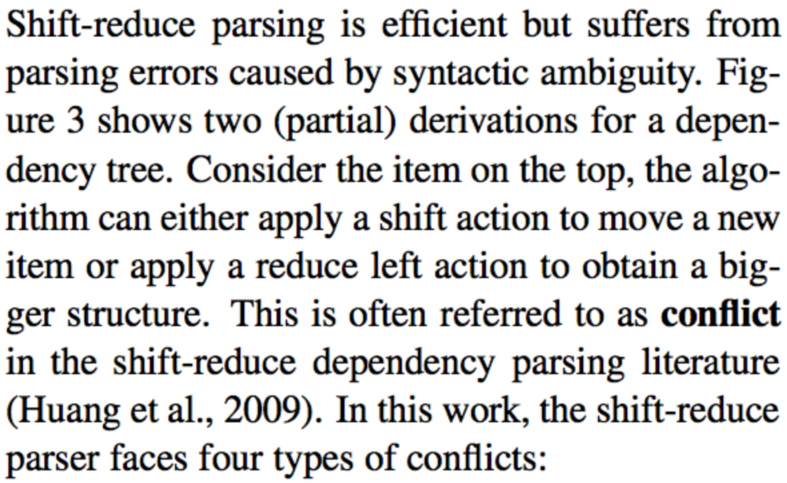
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

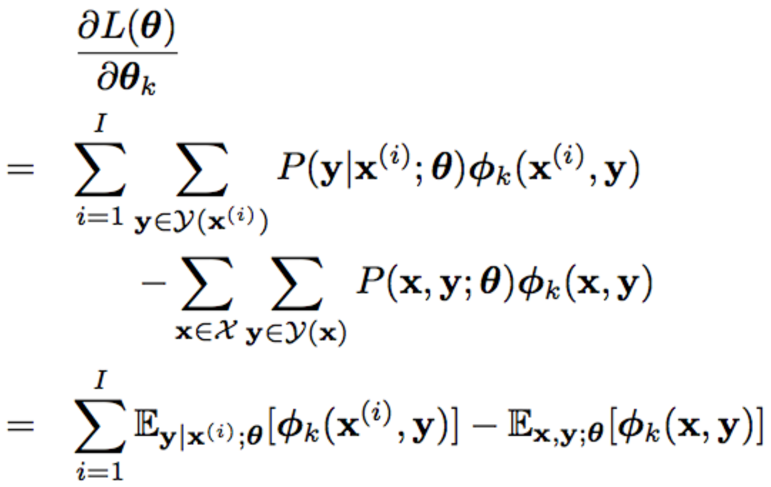
**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

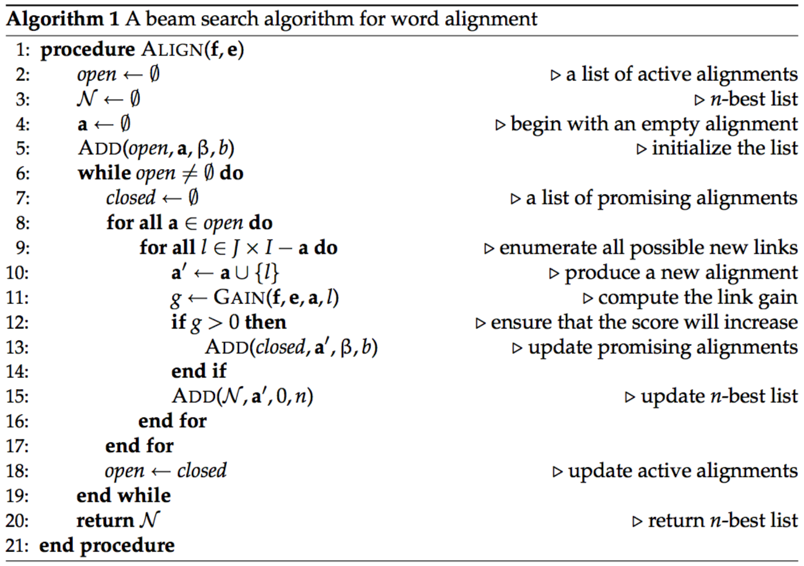
***



$$
\begin{aligned}
& \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} \\
= & \sum_{i=1}^{I} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)}; \boldsymbol{\theta}) \phi_k(\mathbf{x}^{(i)}, \mathbf{y}) \\
& \quad - \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} P(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) \phi_k(\mathbf{x}, \mathbf{y}) \\
= & \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)}, \mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x}, \mathbf{y})]
\end{aligned}
$$

公式

****



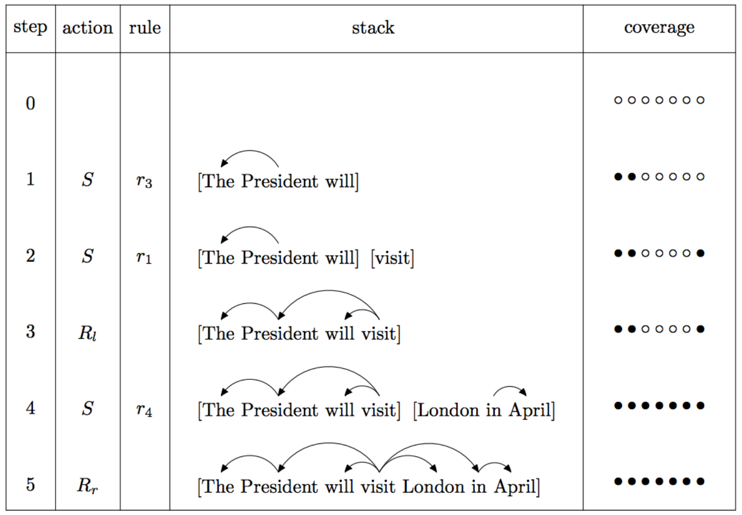
Algorithm 1 A beam search algorithm for word alignment

```
procedure ALIGN(f, e)
    open ← ∅                                  ▷ a list of active alignments
    N ← ∅                                     ▷ n-best list
    a ← ∅                                     ▷ begin with an empty alignment
    ADD(open, a, β, b)                        ▷ initialize the list
    while open ≠ ∅ do
        closed ← ∅                            ▷ a list of promising alignments
        for all a ∈ open do
            for all l ∈ J × I − a do          ▷ enumerate all possible new links
                a′ ← a ∪ {l}                  ▷ produce a new alignment
                g ← GAIN(f, e, a, l)          ▷ compute the link gain
                if g > 0 then                 ▷ ensure that the score will increase
                    ADD(closed, a′, β, b)     ▷ update promising alignments
                end if
                ADD(N, a′, 0, n)              ▷ update n-best list
            end for
        end for
        open ← closed                         ▷ update active alignments
    end while
    return N                                  ▷ return n-best list
end procedure
```

算法

*****

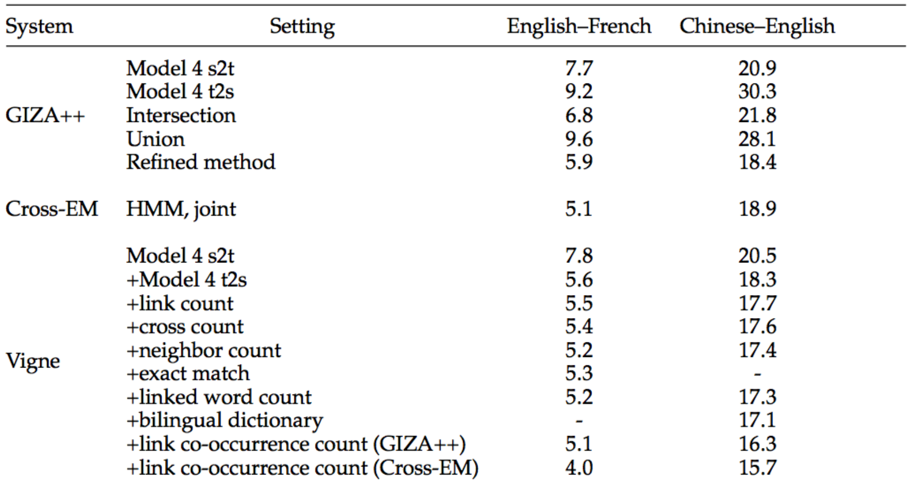
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

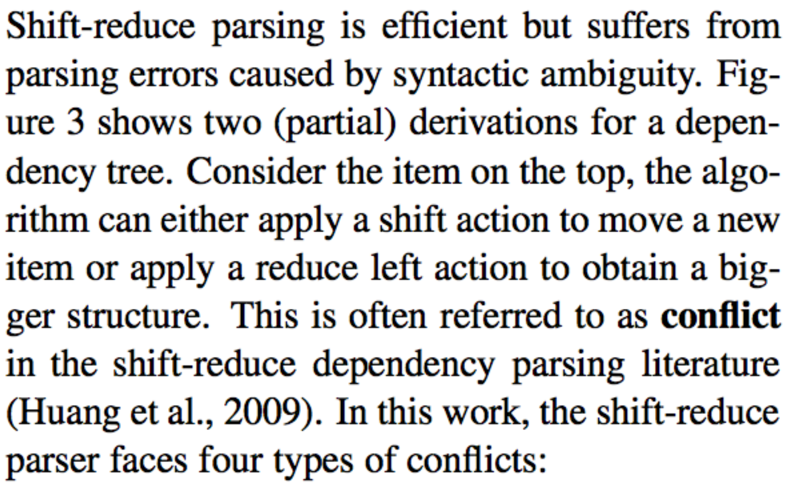
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

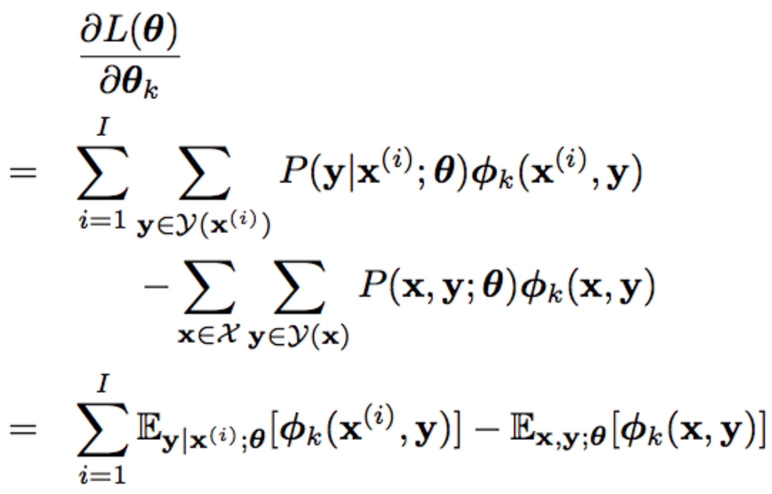
**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

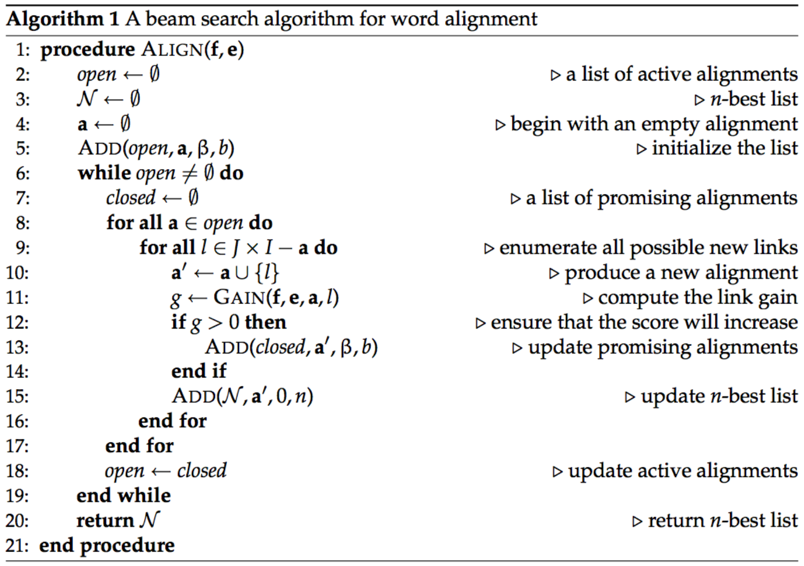
***



$$\begin{aligned}
&\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_k}\\
&= \sum_{i=1}^{I} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta})\phi_k(\mathbf{x}^{(i)},\mathbf{y})\\
&\qquad - \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\phi_k(\mathbf{x},\mathbf{y})\\
&= \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x},\mathbf{y})]
\end{aligned}$$

公式

****

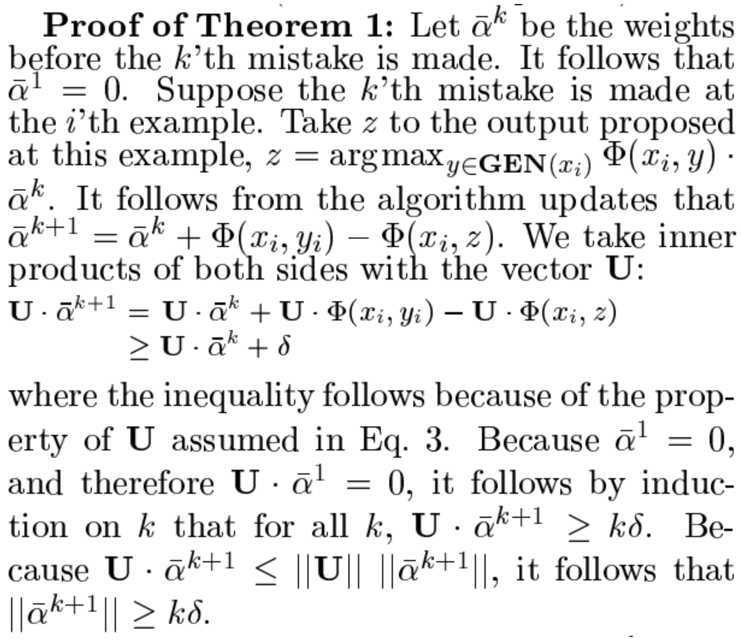


Algorithm 1 A beam search algorithm for word alignment

```
procedure ALIGN($\mathbf{f}, \mathbf{e}$)
    $open \leftarrow \emptyset$                          ▷ a list of active alignments
    $\mathcal{N} \leftarrow \emptyset$                              ▷ $n$-best list
    $\mathbf{a} \leftarrow \emptyset$                      ▷ begin with an empty alignment
    ADD($open, \mathbf{a}, \beta, b$)                      ▷ initialize the list
    while $open \neq \emptyset$ do
        $closed \leftarrow \emptyset$                   ▷ a list of promising alignments
        for all $\mathbf{a} \in open$ do
            for all $l \in J \times I - \mathbf{a}$ do   ▷ enumerate all possible new links
                $\mathbf{a}' \leftarrow \mathbf{a} \cup \{l\}$          ▷ produce a new alignment
                $g \leftarrow$ GAIN($\mathbf{f}, \mathbf{e}, \mathbf{a}, l$)        ▷ compute the link gain
                if $g > 0$ then             ▷ ensure that the score will increase
                    ADD($closed, \mathbf{a}', \beta, b$)   ▷ update promising alignments
                end if
                ADD($\mathcal{N}, \mathbf{a}', 0, n$)                ▷ update $n$-best list
            end for
        end for
        $open \leftarrow closed$                    ▷ update active alignments
    end while
    return $\mathcal{N}$                                  ▷ return $n$-best list
end procedure
```

算法

*****

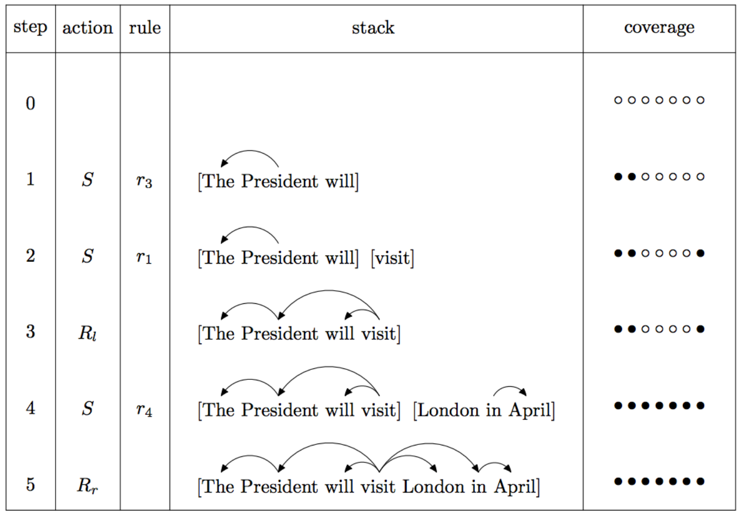


Proof of Theorem 1: Let $\bar{\alpha}^k$ be the weights before the k'th mistake is made. It follows that $\bar{\alpha}^1 = 0$. Suppose the k'th mistake is made at the i'th example. Take z to the output proposed at this example, $z = \arg\max_{y\in\mathbf{GEN}(x_i)} \Phi(x_i, y) \cdot \bar{\alpha}^k$. It follows from the algorithm updates that $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \Phi(x_i, y_i) - \Phi(x_i, z)$. We take inner products of both sides with the vector $\mathbf{U}$:

$$\begin{aligned}
\mathbf{U}\cdot\bar{\alpha}^{k+1} &= \mathbf{U}\cdot\bar{\alpha}^k + \mathbf{U}\cdot\Phi(x_i, y_i) - \mathbf{U}\cdot\Phi(x_i, z)\\
&\geq \mathbf{U}\cdot\bar{\alpha}^k + \delta
\end{aligned}$$

where the inequality follows because of the property of $\mathbf{U}$ assumed in Eq. 3. Because $\bar{\alpha}^1 = 0$, and therefore $\mathbf{U}\cdot\bar{\alpha}^1 = 0$, it follows by induction on k that for all k, $\mathbf{U}\cdot\bar{\alpha}^{k+1} \geq k\delta$. Because $\mathbf{U}\cdot\bar{\alpha}^{k+1} \leq ||\mathbf{U}||\ ||\bar{\alpha}^{k+1}||$, it follows that $||\bar{\alpha}^{k+1}|| \geq k\delta$.

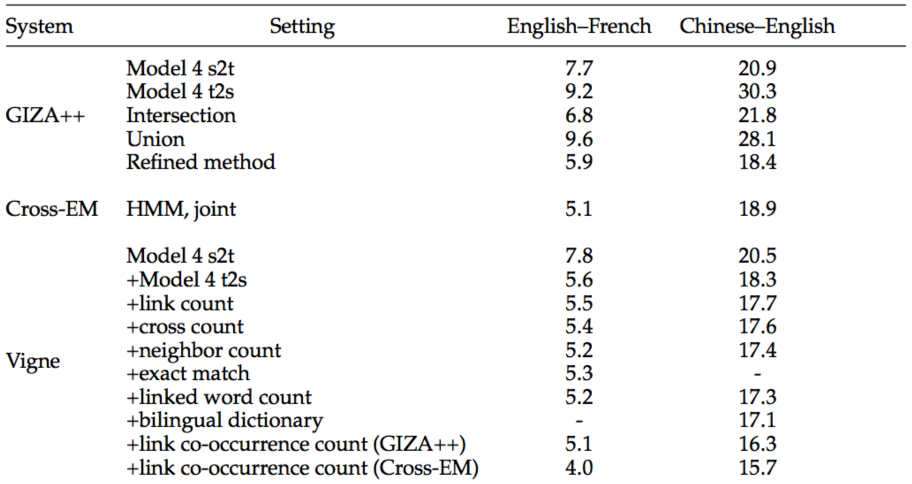
信息元素的易理解度



step	action	rule	stack	coverage
0				○○○○○○○
1	S	r_3	[The President will]	●●○○○○○
2	S	r_1	[The President will] [visit]	●●○○○○●
3	R_l		[The President will visit]	●●○○○○●
4	S	r_4	[The President will visit] [London in April]	●●●●●●●
5	R_r		[The President will visit London in April]	●●●●●●●

图

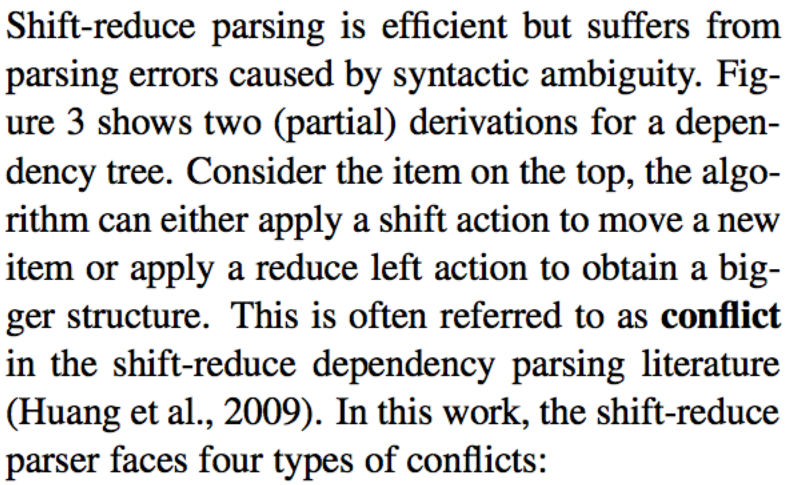
*



System	Setting	English–French	Chinese–English
GIZA++	Model 4 s2t	7.7	20.9
	Model 4 t2s	9.2	30.3
	Intersection	6.8	21.8
	Union	9.6	28.1
	Refined method	5.9	18.4
Cross-EM	HMM, joint	5.1	18.9
Vigne	Model 4 s2t	7.8	20.5
	+Model 4 t2s	5.6	18.3
	+link count	5.5	17.7
	+cross count	5.4	17.6
	+neighbor count	5.2	17.4
	+exact match	5.3	-
	+linked word count	5.2	17.3
	+bilingual dictionary	-	17.1
	+link co-occurrence count (GIZA++)	5.1	16.3
	+link co-occurrence count (Cross-EM)	4.0	15.7

表格

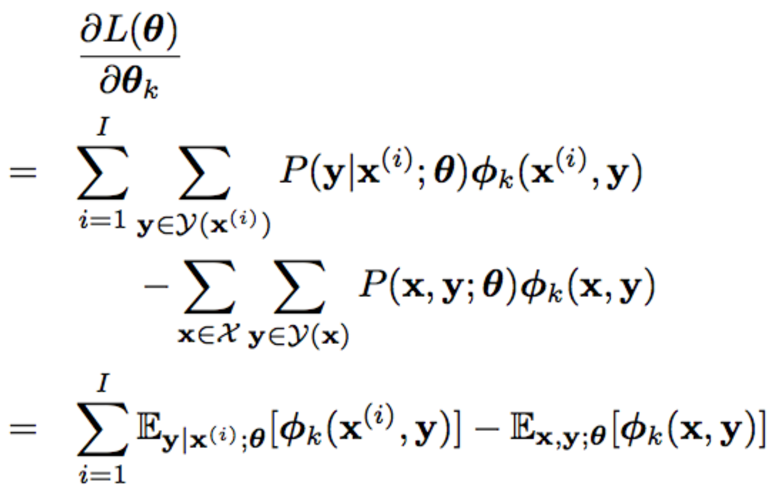
**



Shift-reduce parsing is efficient but suffers from parsing errors caused by syntactic ambiguity. Figure 3 shows two (partial) derivations for a dependency tree. Consider the item on the top, the algorithm can either apply a shift action to move a new item or apply a reduce left action to obtain a bigger structure. This is often referred to as **conflict** in the shift-reduce dependency parsing literature (Huang et al., 2009). In this work, the shift-reduce parser faces four types of conflicts:

正文

***



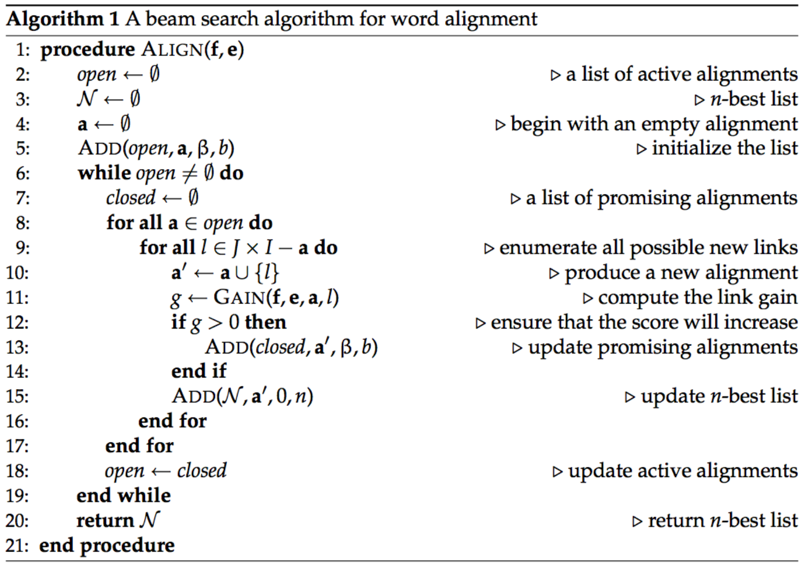
$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k}$$

$$= \sum_{i=1}^{I} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x}^{(i)})} P(\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta})\phi_k(\mathbf{x}^{(i)},\mathbf{y}) - \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} P(\mathbf{x},\mathbf{y};\boldsymbol{\theta})\phi_k(\mathbf{x},\mathbf{y})$$

$$= \sum_{i=1}^{I} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(i)};\boldsymbol{\theta}}[\phi_k(\mathbf{x}^{(i)},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y};\boldsymbol{\theta}}[\phi_k(\mathbf{x},\mathbf{y})]$$

公式

****

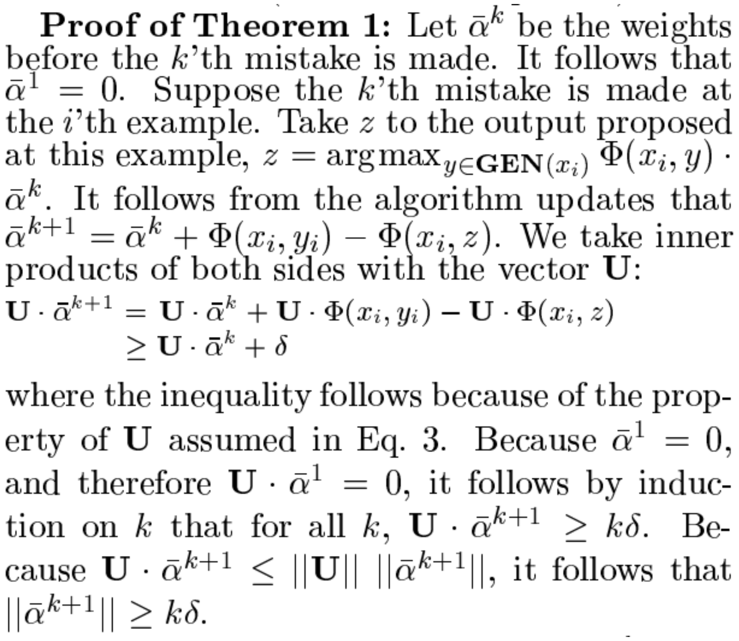


Algorithm 1 A beam search algorithm for word alignment

```
procedure ALIGN($\mathbf{f}, \mathbf{e}$)
    $open \leftarrow \emptyset$                          ▷ a list of active alignments
    $\mathcal{N} \leftarrow \emptyset$                          ▷ n-best list
    $\mathbf{a} \leftarrow \emptyset$                            ▷ begin with an empty alignment
    ADD($open, \mathbf{a}, \beta, b$)                  ▷ initialize the list
    while $open \neq \emptyset$ do
        $closed \leftarrow \emptyset$                    ▷ a list of promising alignments
        for all $\mathbf{a} \in open$ do
            for all $l \in J \times I - \mathbf{a}$ do      ▷ enumerate all possible new links
                $\mathbf{a}' \leftarrow \mathbf{a} \cup \{l\}$            ▷ produce a new alignment
                $g \leftarrow$ GAIN($\mathbf{f}, \mathbf{e}, \mathbf{a}, l$)        ▷ compute the link gain
                if $g > 0$ then                  ▷ ensure that the score will increase
                    ADD($closed, \mathbf{a}', \beta, b$)      ▷ update promising alignments
                end if
                ADD($\mathcal{N}, \mathbf{a}', 0, n$)          ▷ update n-best list
            end for
        end for
        $open \leftarrow closed$                 ▷ update active alignments
    end while
    return $\mathcal{N}$                            ▷ return n-best list
end procedure
```

算法

*****



Proof of Theorem 1: Let $\bar{\alpha}^k$ be the weights before the k'th mistake is made. It follows that $\bar{\alpha}^1 = 0$. Suppose the k'th mistake is made at the i'th example. Take z to the output proposed at this example, $z = \arg\max_{y\in\mathbf{GEN}(x_i)} \Phi(x_i, y) \cdot \bar{\alpha}^k$. It follows from the algorithm updates that $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \Phi(x_i, y_i) - \Phi(x_i, z)$. We take inner products of both sides with the vector $\mathbf{U}$:

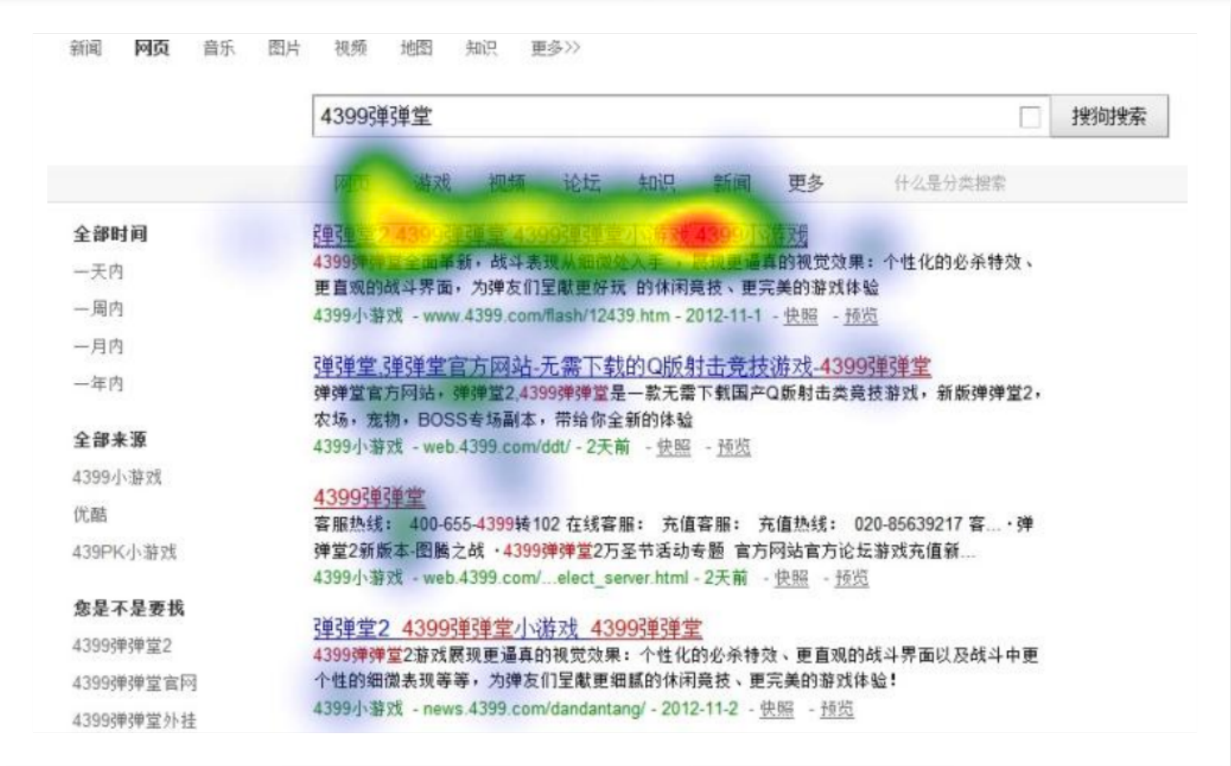
$$\begin{aligned}\mathbf{U}\cdot\bar{\alpha}^{k+1} &= \mathbf{U}\cdot\bar{\alpha}^k + \mathbf{U}\cdot\Phi(x_i, y_i) - \mathbf{U}\cdot\Phi(x_i, z)\\ &\geq \mathbf{U}\cdot\bar{\alpha}^k + \delta\end{aligned}$$

where the inequality follows because of the property of $\mathbf{U}$ assumed in Eq. 3. Because $\bar{\alpha}^1 = 0$, and therefore $\mathbf{U}\cdot\bar{\alpha}^1 = 0$, it follows by induction on k that for all k, $\mathbf{U}\cdot\bar{\alpha}^{k+1} \geq k\delta$. Because $\mathbf{U}\cdot\bar{\alpha}^{k+1} \leq ||\mathbf{U}||\ ||\bar{\alpha}^{k+1}||$, it follows that $||\bar{\alpha}^{k+1}|| \geq k\delta$.

证明

******

眼动仪的佐证



弹弹堂2 4399弹弹堂 4399弹弹堂小游戏 4399小游戏
4399弹弹堂全面革新，战斗表现从细微处入手，展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面，为弹友们呈献更好玩 的休闲竞技、更完美的游戏体验
4399小游戏 - www.4399.com/flash/12439.htm - 2012-11-1 - 快照 - 预览

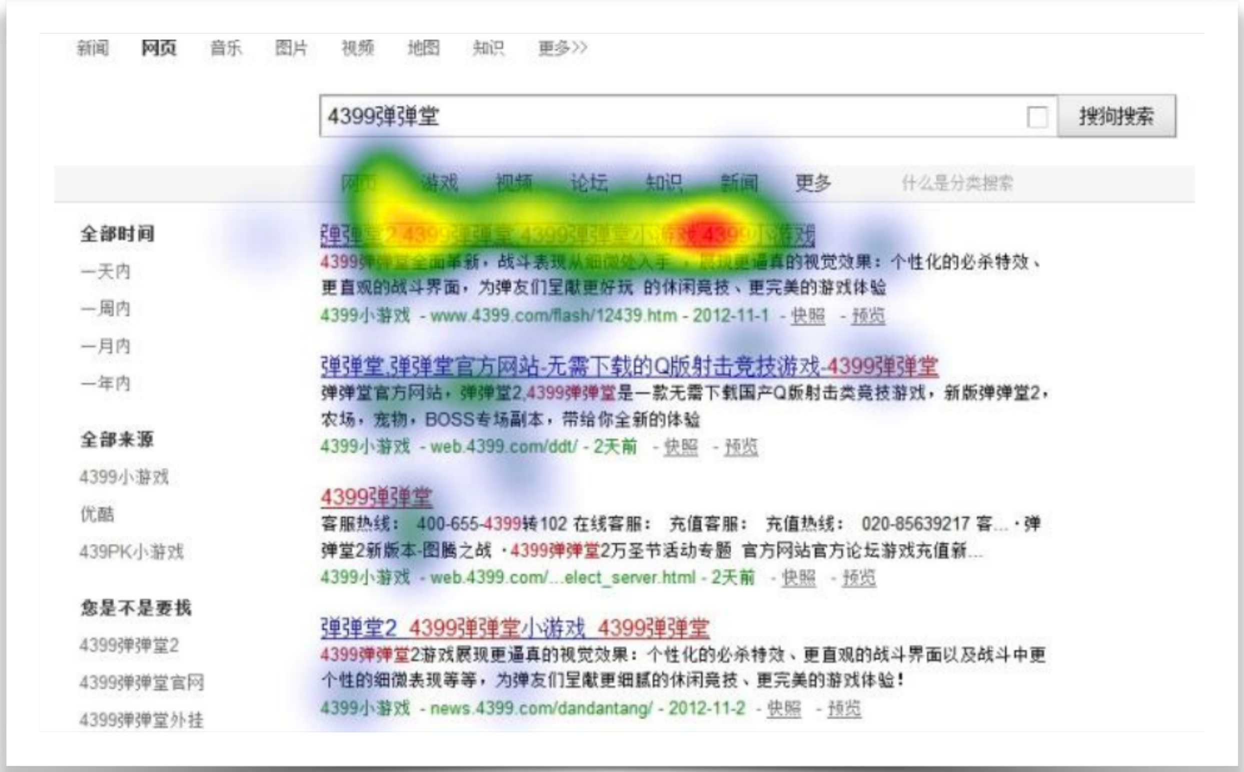
弹弹堂,弹弹堂官方网站-无需下载的Q版射击竞技游戏-4399弹弹堂
弹弹堂官方网站，弹弹堂2,4399弹弹堂是一款无需下载国产Q版射击类竞技游戏，新版弹弹堂2，农场，宠物，BOSS专场副本，带给你全新的体验
4399小游戏 - web.4399.com/ddt/ - 2天前 - 快照 - 预览

4399弹弹堂
客服热线： 400-655-4399转102 在线客服： 充值客服： 充值热线： 020-85639217 客... ·弹弹堂2新版本-图腾之战 ·4399弹弹堂2万圣节活动专题 官方网站官方论坛游戏充值新...
4399小游戏 - web.4399.com/...elect_server.html - 2天前 - 快照 - 预览

弹弹堂2 4399弹弹堂小游戏 4399弹弹堂
4399弹弹堂2游戏展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面以及战斗中更个性的细微表现等等，为弹友们呈献更细腻的休闲竞技、更完美的游戏体验！
4399小游戏 - news.4399.com/dandantang/ - 2012-11-2 - 快照 - 预览

图片来自清华大学刘奕群

眼动仪的佐证

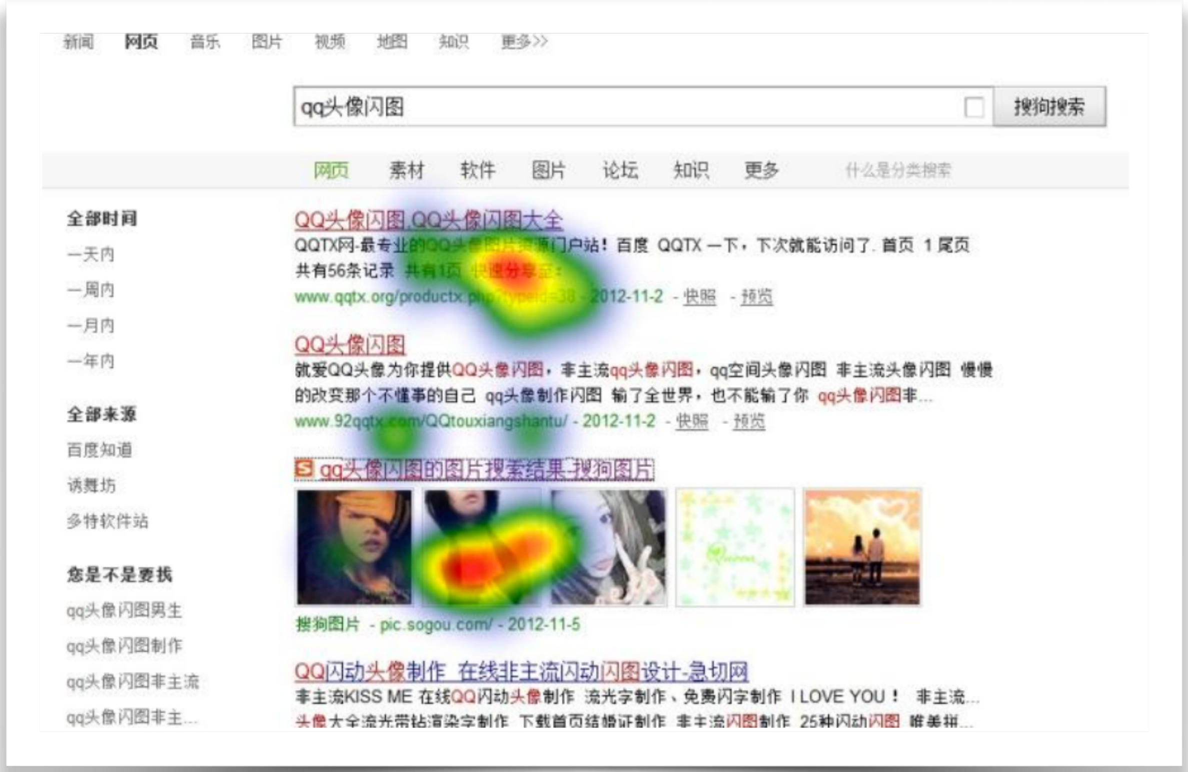


弹弹堂2 4399弹弹堂 4399弹弹堂小游戏 4399小游戏
4399弹弹堂全面革新，战斗表现从细微处入手，展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面，为弹友们呈献更好玩 的休闲竞技、更完美的游戏体验
4399小游戏 - www.4399.com/flash/12439.htm - 2012-11-1 - 快照 - 预览

弹弹堂,弹弹堂官方网站-无需下载的Q版射击竞技游戏-4399弹弹堂
弹弹堂官方网站，弹弹堂2,4399弹弹堂是一款无需下载国产Q版射击类竞技游戏，新版弹弹堂2，农场，宠物，BOSS专场副本，带给你全新的体验
4399小游戏 - web.4399.com/ddt/ - 2天前 - 快照 - 预览

4399弹弹堂
客服热线： 400-655-4399转102 在线客服： 充值客服： 充值热线： 020-85639217 客... ·弹弹堂2新版本-图腾之战 ·4399弹弹堂2万圣节活动专题 官方网站官方论坛游戏充值新...
4399小游戏 - web.4399.com/...elect_server.html - 2天前 - 快照 - 预览

弹弹堂2 4399弹弹堂小游戏 4399弹弹堂
4399弹弹堂2游戏展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面以及战斗中更个性的细微表现等等，为弹友们呈献更细腻的休闲竞技、更完美的游戏体验！
4399小游戏 - news.4399.com/dandantang/ - 2012-11-2 - 快照 - 预览



QQ头像闪图,QQ头像闪图大全
QQTX网-最专业的QQ头像图片资源门户站！百度 QQTX 一下，下次就能访问了. 首页 1 尾页 共有56条记录 共有1页 快速分享至：
www.qqtx.org/productx.php?typeid=38 - 2012-11-2 - 快照 - 预览

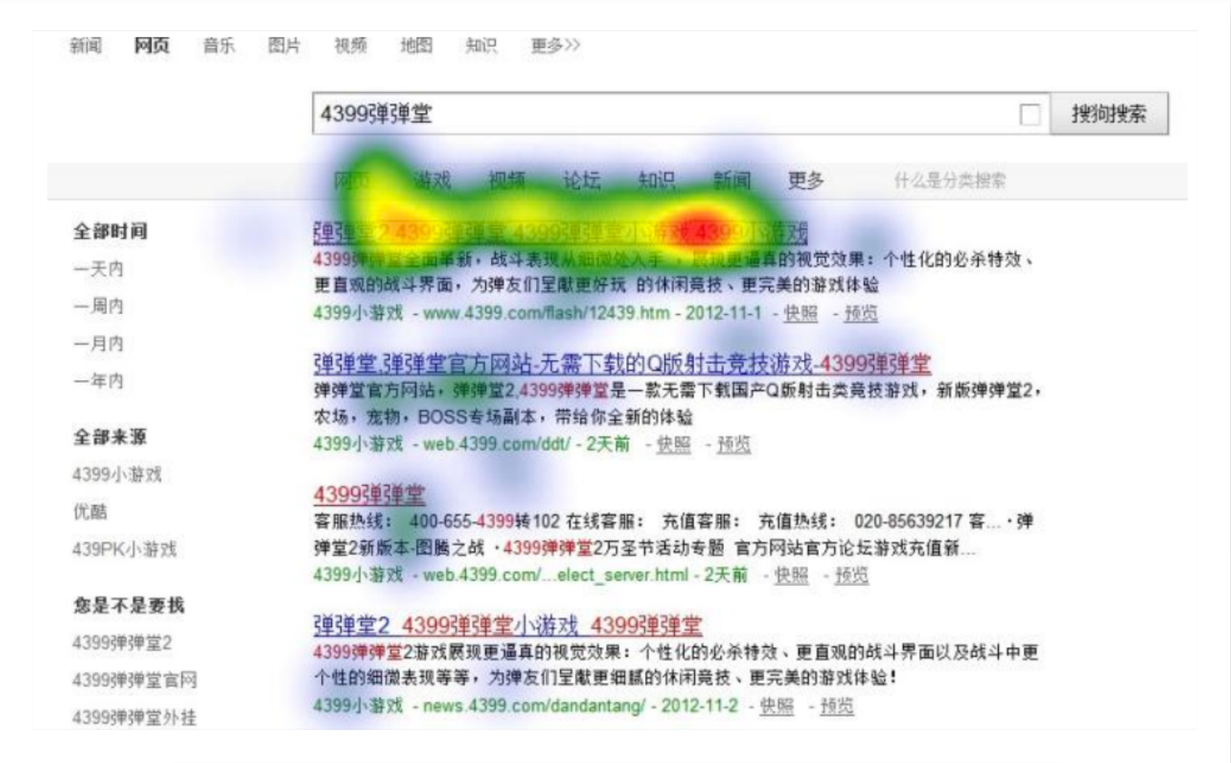
QQ头像闪图
就爱QQ头像为你提供QQ头像闪图，非主流qq头像闪图，qq空间头像闪图 非主流头像闪图 慢慢的改变那个不懂事的自己 qq头像制作闪图 输了全世界，也不能输了你 qq头像闪图非...
www.92qqtx.com/QQtouxiangshantu/ - 2012-11-2 - 快照 - 预览

S qq头像闪图的图片搜索结果-搜狗图片
搜狗图片 - pic.sogou.com/ - 2012-11-5

QQ闪动头像制作 在线非主流闪动闪图设计-急切网
非主流KISS ME 在线QQ闪动头像制作 流光字制作、免费闪字制作 I LOVE YOU！ 非主流...
头像大全流光带钻渲染字制作 下载首页结婚证制作 非主流闪图制作 25种闪动闪图 唯美拼...

图片来自清华大学刘奕群

眼动仪的佐证

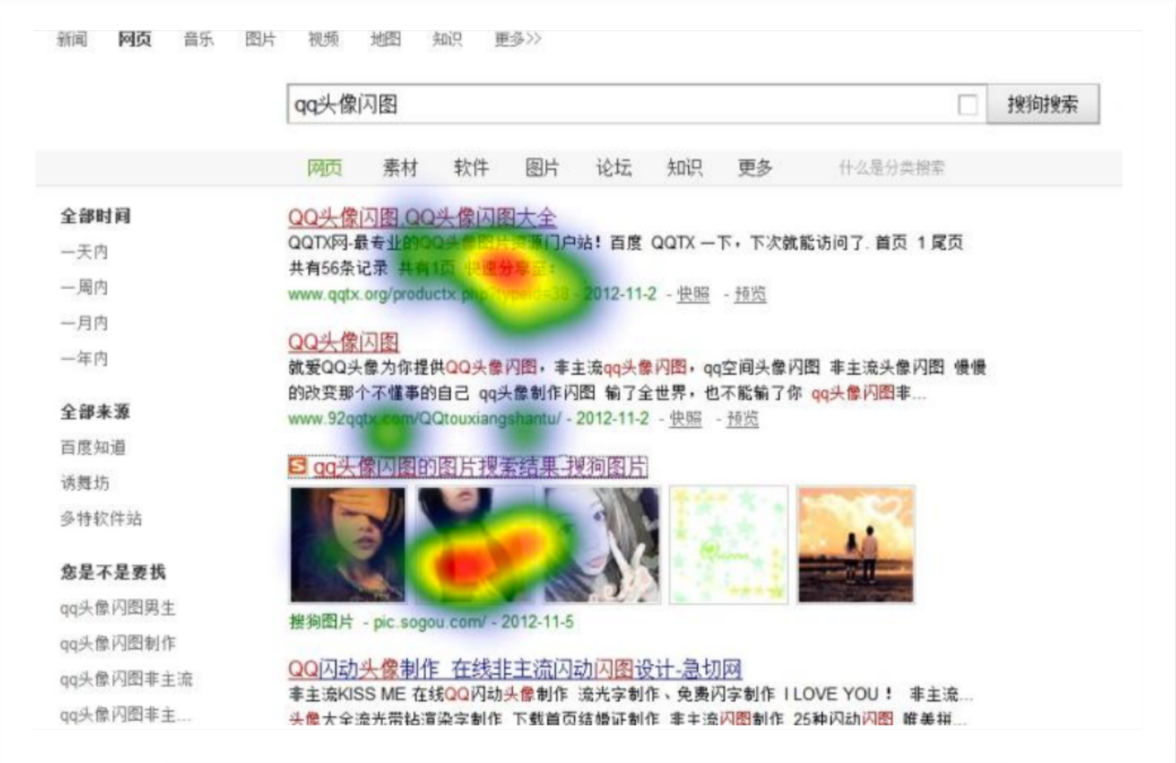


弹弹堂2 4399弹弹堂 4399弹弹堂小游戏 4399小游戏
4399弹弹堂全面革新，战斗表现从细微处入手，展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面，为弹友们呈献更好玩 的休闲竞技、更完美的游戏体验
4399小游戏 - www.4399.com/flash/12439.htm - 2012-11-1 - 快照 - 预览

弹弹堂,弹弹堂官方网站-无需下载的Q版射击竞技游戏-4399弹弹堂
弹弹堂官方网站，弹弹堂2,4399弹弹堂是一款无需下载国产Q版射击类竞技游戏，新版弹弹堂2，农场，宠物，BOSS专场副本，带给你全新的体验
4399小游戏 - web.4399.com/ddt/ - 2天前 - 快照 - 预览

4399弹弹堂
客服热线： 400-655-4399转102 在线客服： 充值客服： 充值热线： 020-85639217 客... ·弹弹堂2新版本-图腾之战 ·4399弹弹堂2万圣节活动专题 官方网站官方论坛游戏充值新...
4399小游戏 - web.4399.com/...elect_server.html - 2天前 - 快照 - 预览

弹弹堂2 4399弹弹堂小游戏 4399弹弹堂
4399弹弹堂2游戏展现更逼真的视觉效果：个性化的必杀特效、更直观的战斗界面以及战斗中更个性的细微表现等等，为弹友们呈献更细腻的休闲竞技、更完美的游戏体验！
4399小游戏 - news.4399.com/dandantang/ - 2012-11-2 - 快照 - 预览



QQ头像闪图,QQ头像闪图大全
QQTX网-最专业的QQ头像图片资源门户站！百度 QQTX 一下，下次就能访问了. 首页 1 尾页 共有56条记录 共有1页 快速分享至：
www.qqtx.org/productx.php?typeid=38 - 2012-11-2 - 快照 - 预览

QQ头像闪图
就爱QQ头像为你提供QQ头像闪图，非主流qq头像闪图，qq空间头像闪图 非主流头像闪图 慢慢的改变那个不懂事的自己 qq头像制作闪图 输了全世界，也不能输了你 qq头像闪图非...
www.92qqtx.com/QQtouxiangshantu/ - 2012-11-2 - 快照 - 预览

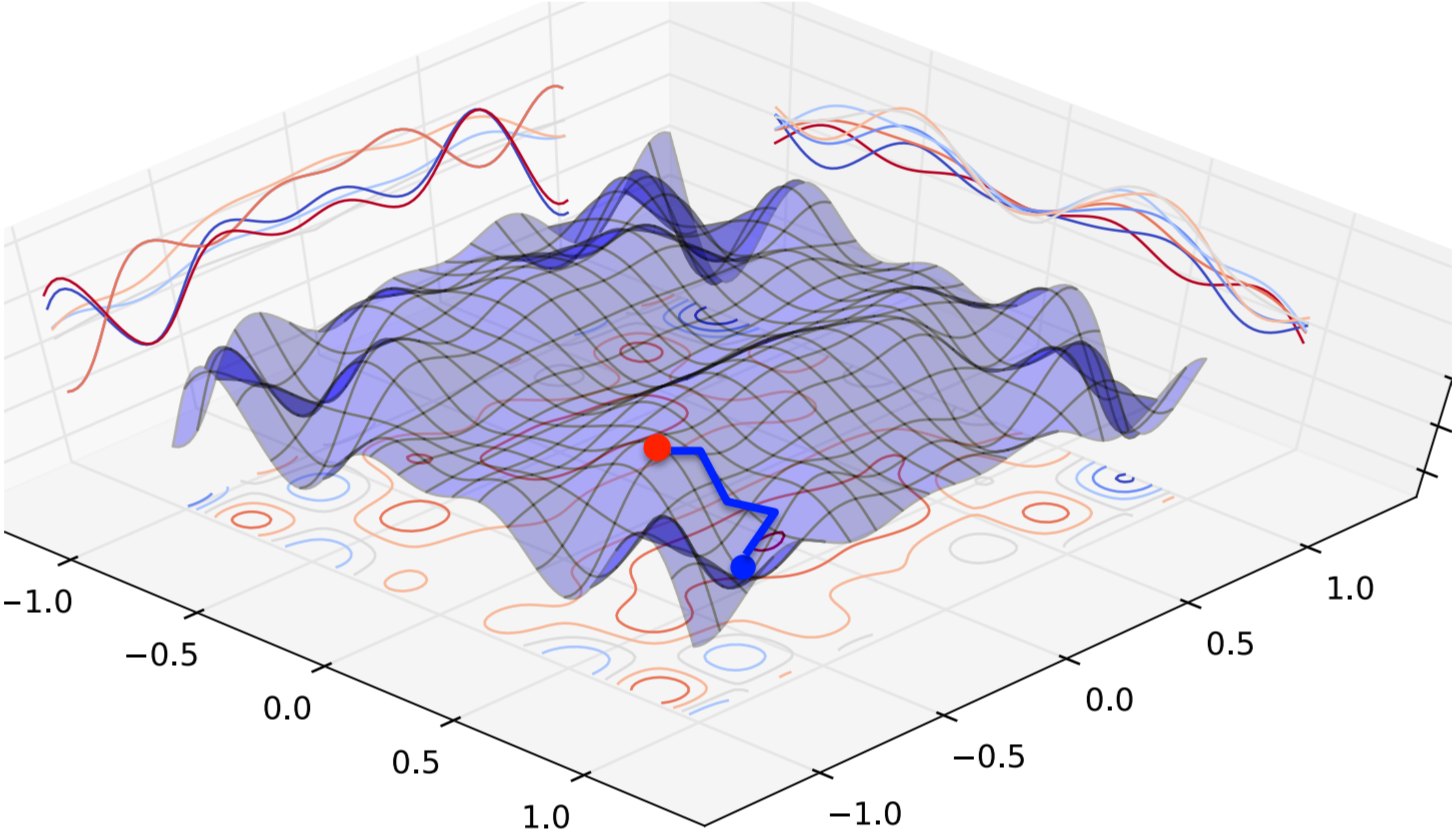
S qq头像闪图的图片搜索结果-搜狗图片
搜狗图片 - pic.sogou.com/ - 2012-11-5

QQ闪动头像制作 在线非主流闪动闪图设计-急切网
非主流KISS ME 在线QQ闪动头像制作 流光字制作、免费闪字制作 I LOVE YOU！ 非主流...
头像大全流光带钻渲染字制作 下载首页结婚证制作 非主流闪图制作 25种闪动闪图 唯美拼...

图片来自清华大学刘奕群

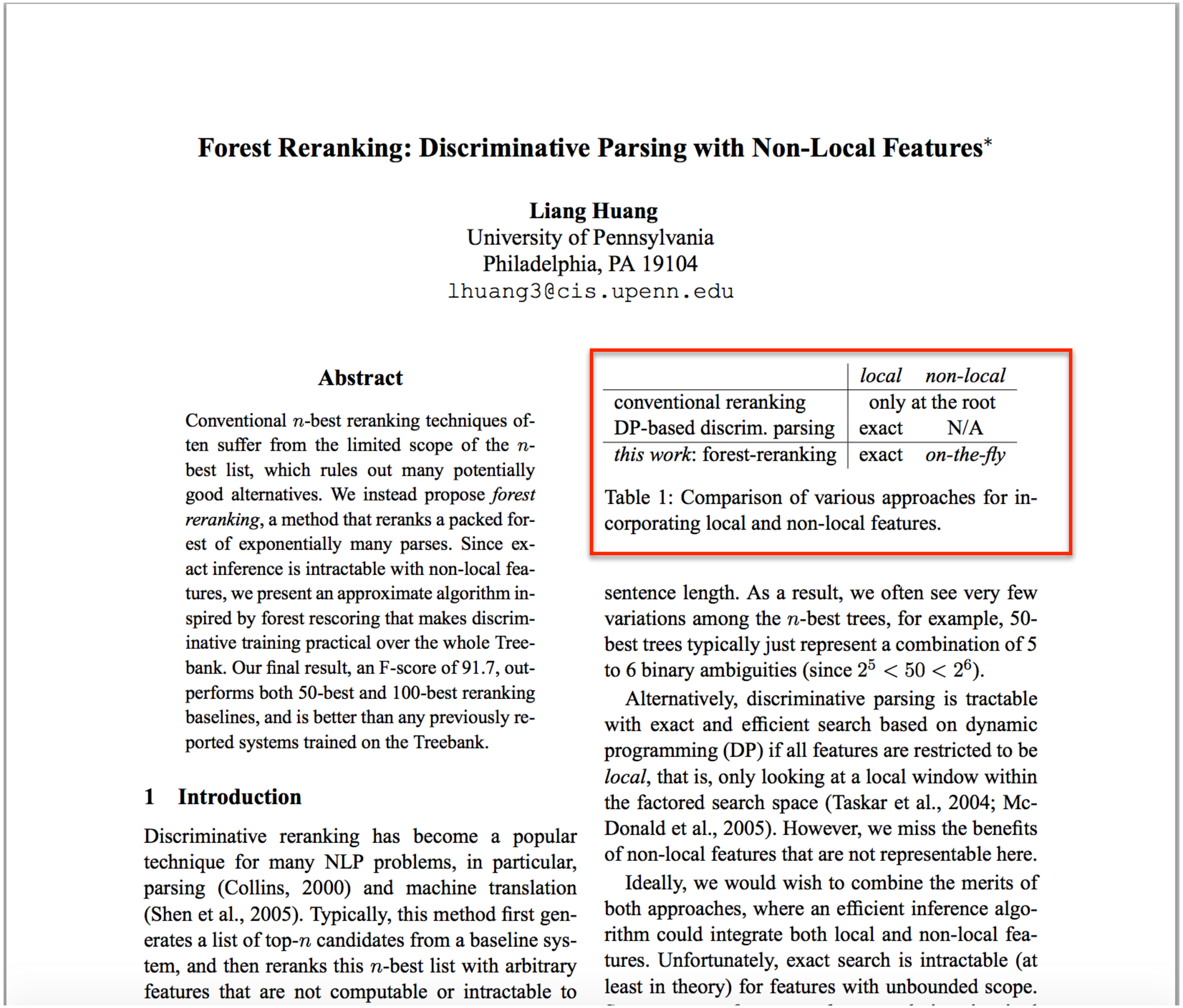
读者潜意识里优先选择易理解度高的信息元素

精心布局，引人入胜



精心对信息元素进行布局，
引导读者在接收信息时走一条“舒服”的路径

小技巧：首页加图表



Forest Reranking: Discriminative Parsing with Non-Local Features*

Liang Huang
University of Pennsylvania
Philadelphia, PA 19104
lhuang3@cis.upenn.edu

Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

1 Introduction

Discriminative reranking has become a popular technique for many NLP problems, in particular, parsing (Collins, 2000) and machine translation (Shen et al., 2005). Typically, this method first generates a list of top-n candidates from a baseline system, and then reranks this n-best list with arbitrary features that are not computable or intractable to

	local	*non-local*
conventional reranking	only at the root	
DP-based discrim. parsing	exact	N/A
this work: forest-reranking	exact	*on-the-fly*

Table 1: Comparison of various approaches for incorporating local and non-local features.

sentence length. As a result, we often see very few variations among the n-best trees, for example, 50-best trees typically just represent a combination of 5 to 6 binary ambiguities (since $2^5 < 50 < 2^6$).

Alternatively, discriminative parsing is tractable with exact and efficient search based on dynamic programming (DP) if all features are restricted to be *local*, that is, only looking at a local window within the factored search space (Taskar et al., 2004; McDonald et al., 2005). However, we miss the benefits of non-local features that are not representable here.

Ideally, we would wish to combine the merits of both approaches, where an efficient inference algorithm could integrate both local and non-local features. Unfortunately, exact search is intractable (at least in theory) for features with unbounded scope.

Liang Huang. **Forest Reranking: Discriminative Parsing with Non-Local Features**. In *ACL 2008*.

信息流的变化

Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , **Qun Liu** , and **Shouxun Lin**
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

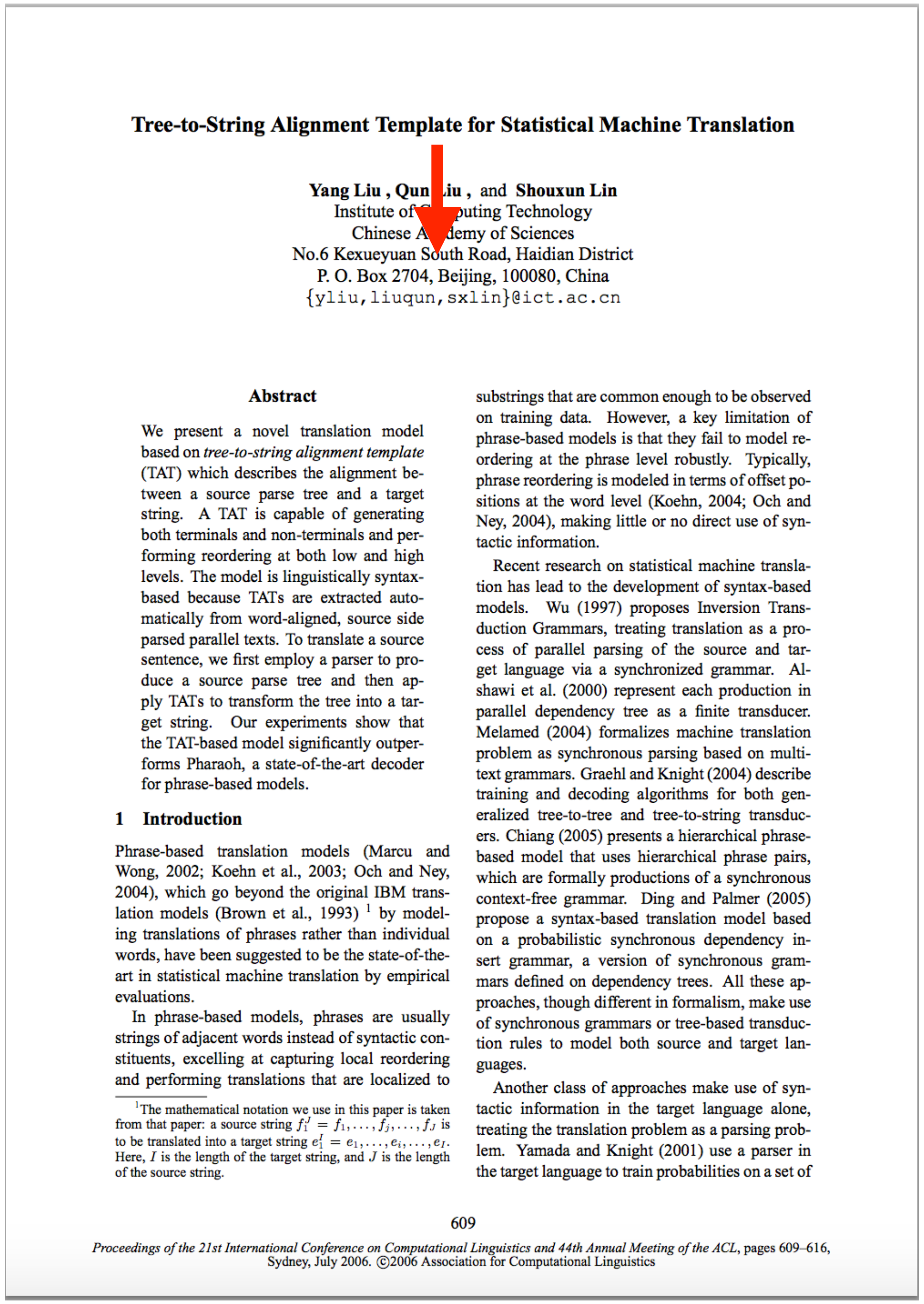
Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

609

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

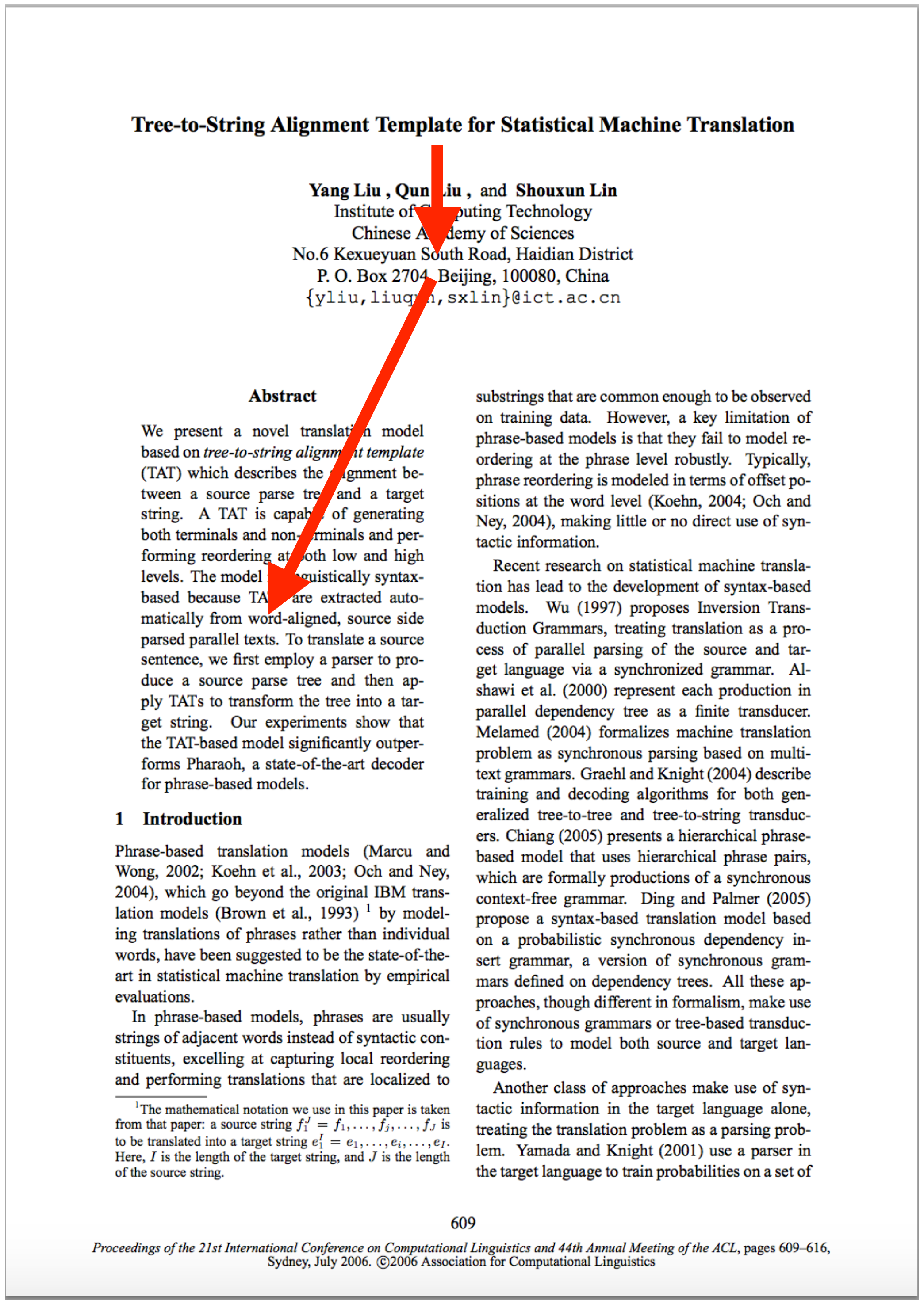
Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

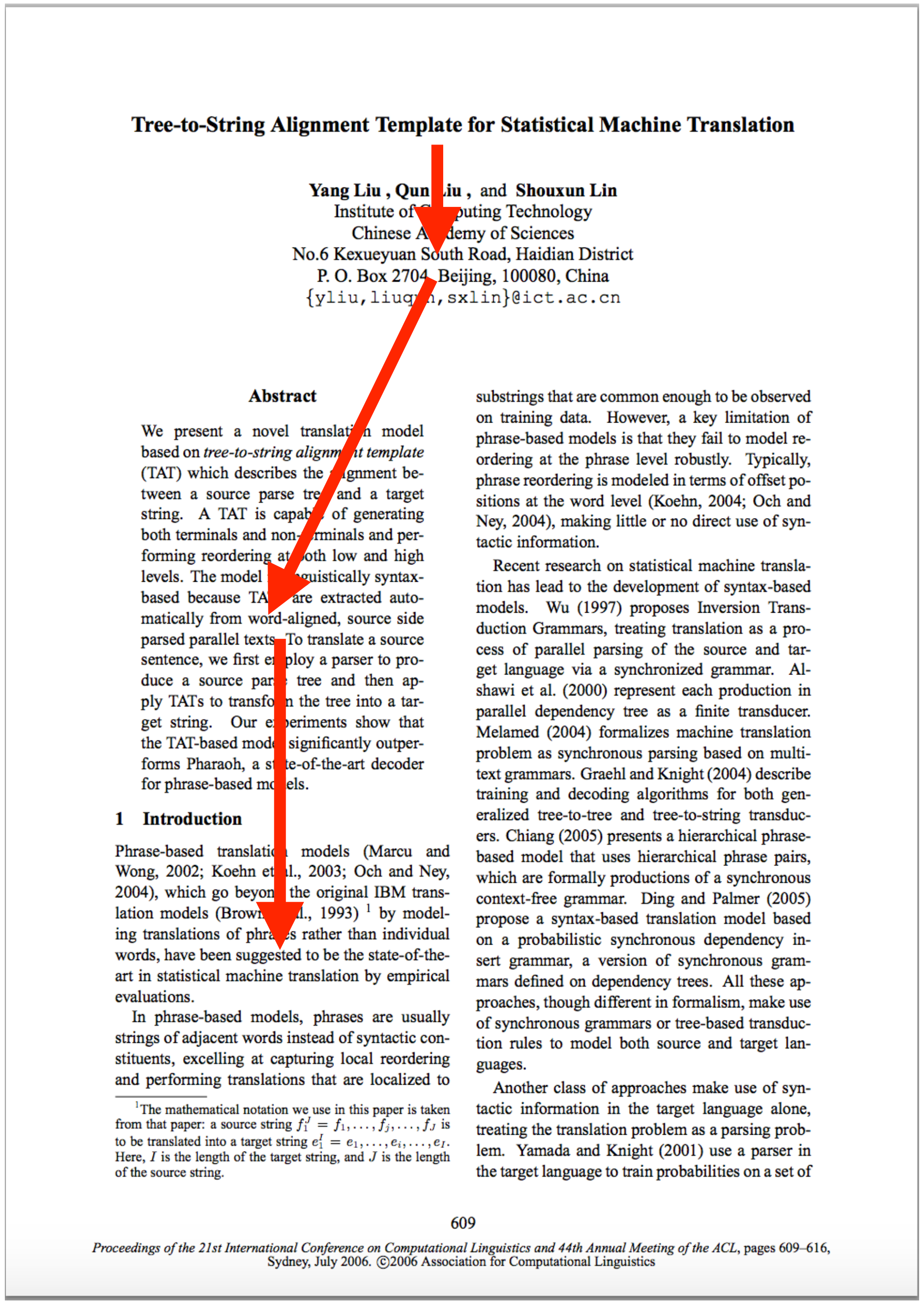
Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \dots, f_j, \dots, f_J$ is to be translated into a target string $e_1^I = e_1, \dots, e_i, \dots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

609

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , **Qun Liu** , and **Shouxun Lin**
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

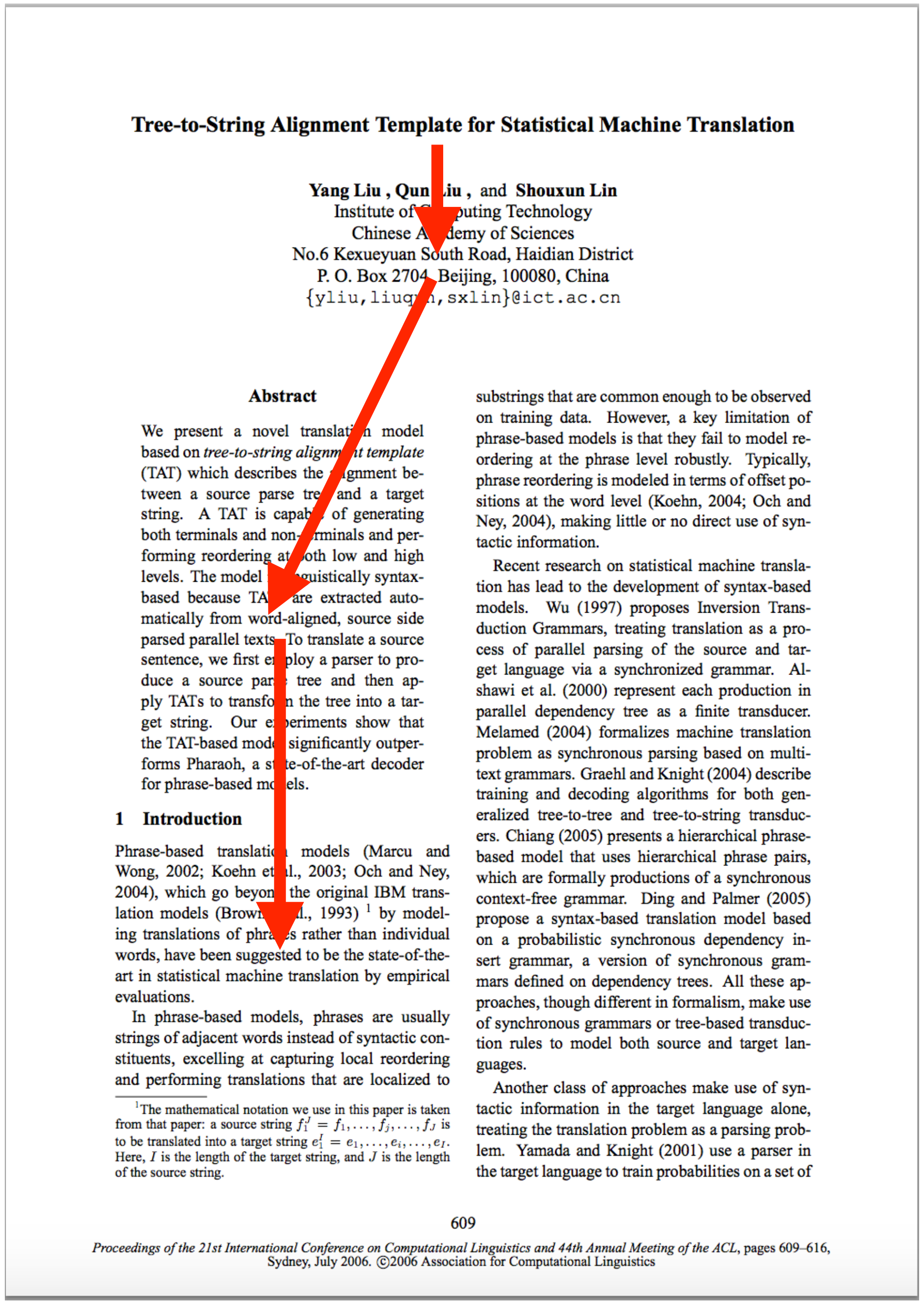
Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

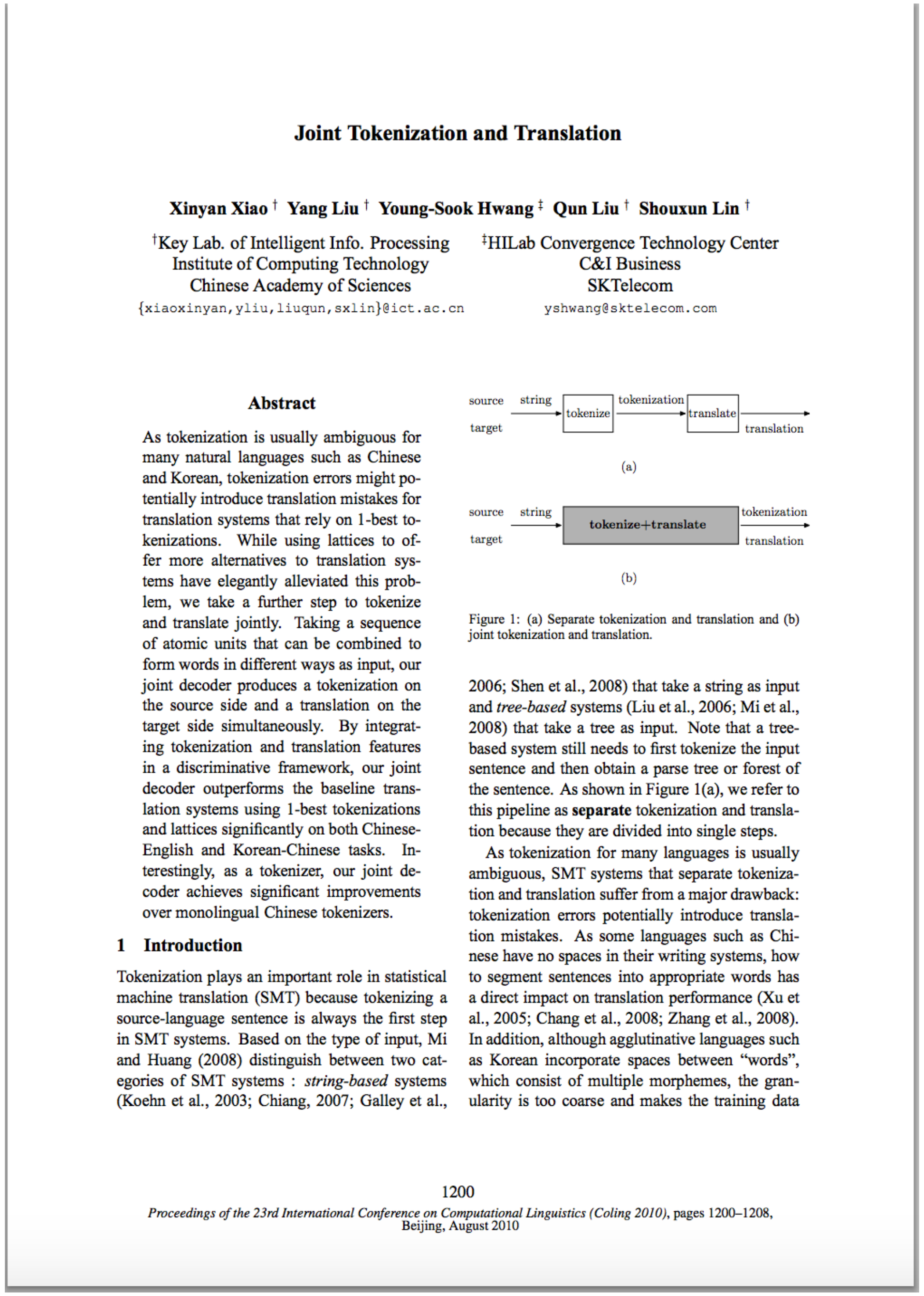
Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

609

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics



Joint Tokenization and Translation

Xinyan Xiao [†] Yang Liu [†] Young-Sook Hwang [‡] Qun Liu [†] Shouxun Lin [†]

[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn

[‡]HILab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a sequence of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

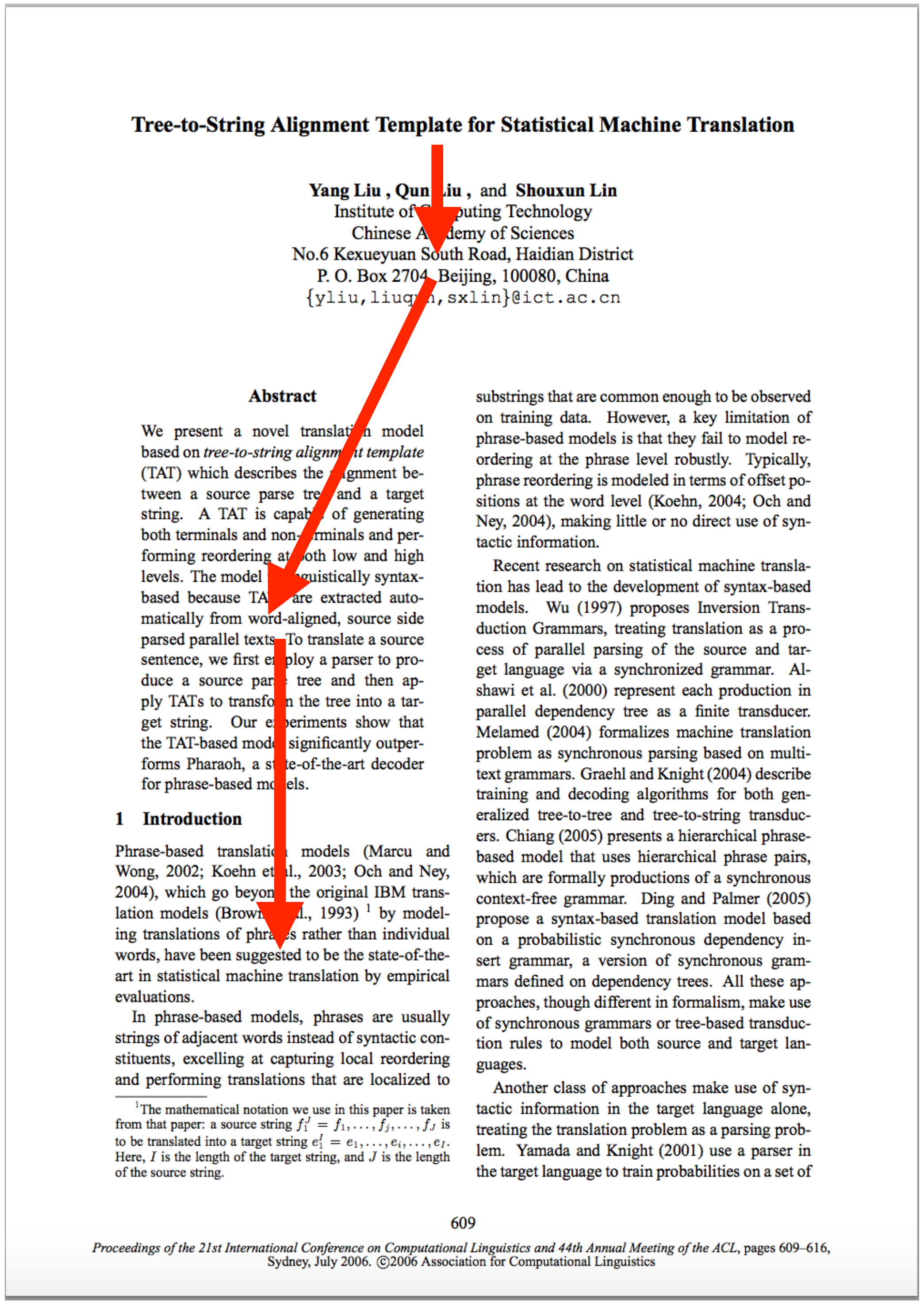
As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

1200

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208,
Beijing, August 2010

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

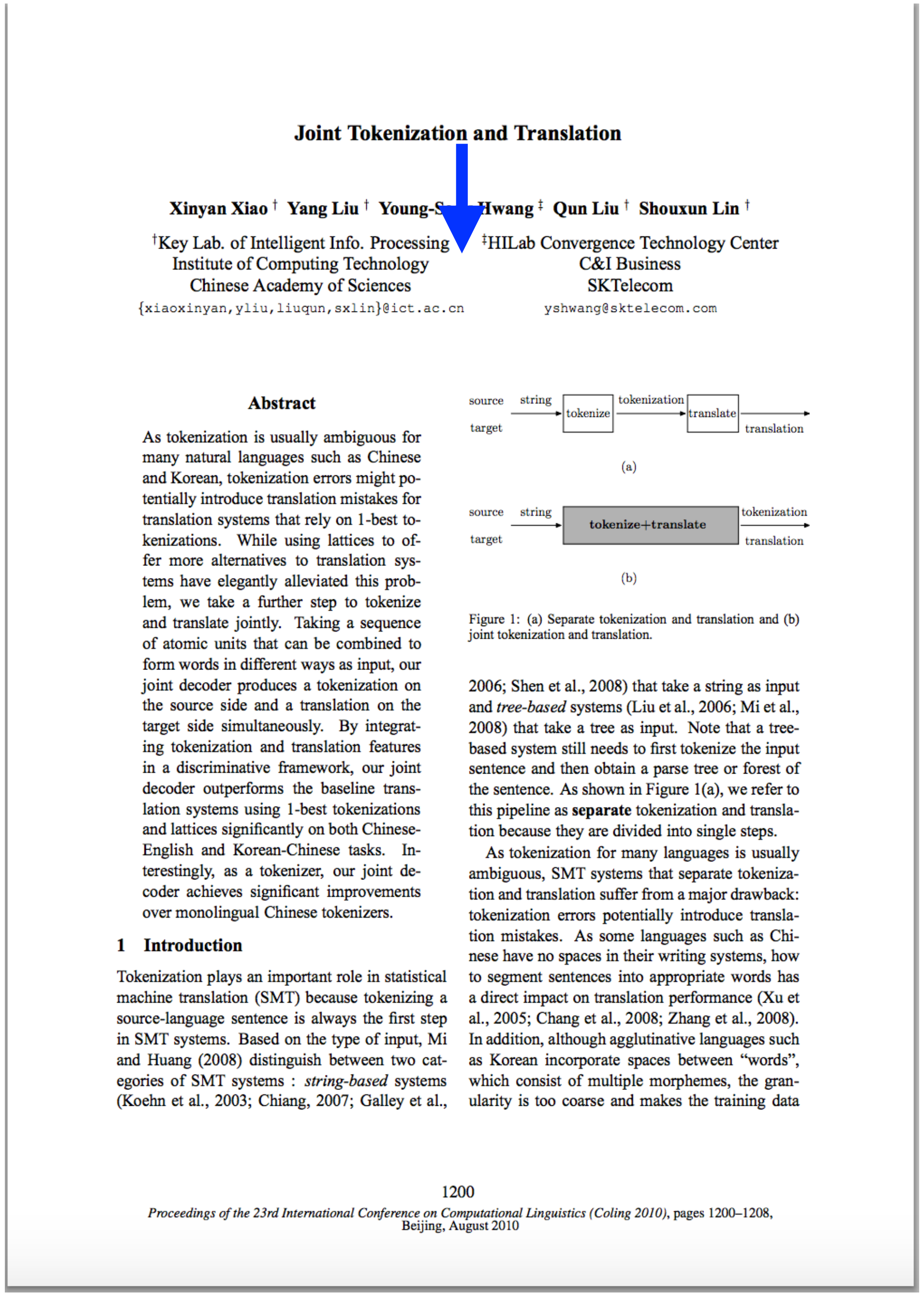
In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics



Joint Tokenization and Translation

Xinyan Xiao [†] **Yang Liu** [†] **Young-Sook Hwang** [‡] **Qun Liu** [†] **Shouxun Lin** [†]

[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn

[‡]HILab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a sequence of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

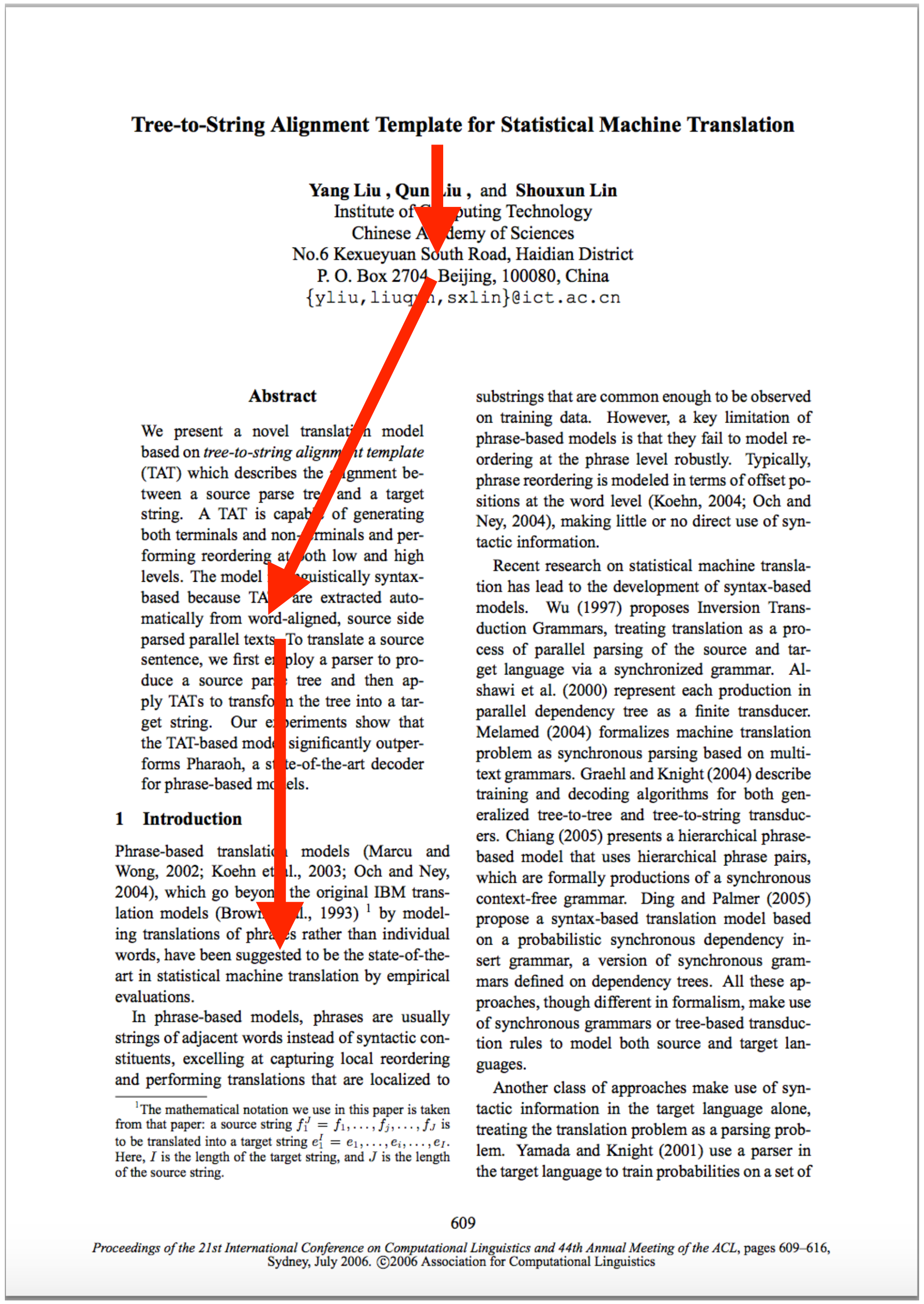
Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208,
Beijing, August 2010

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

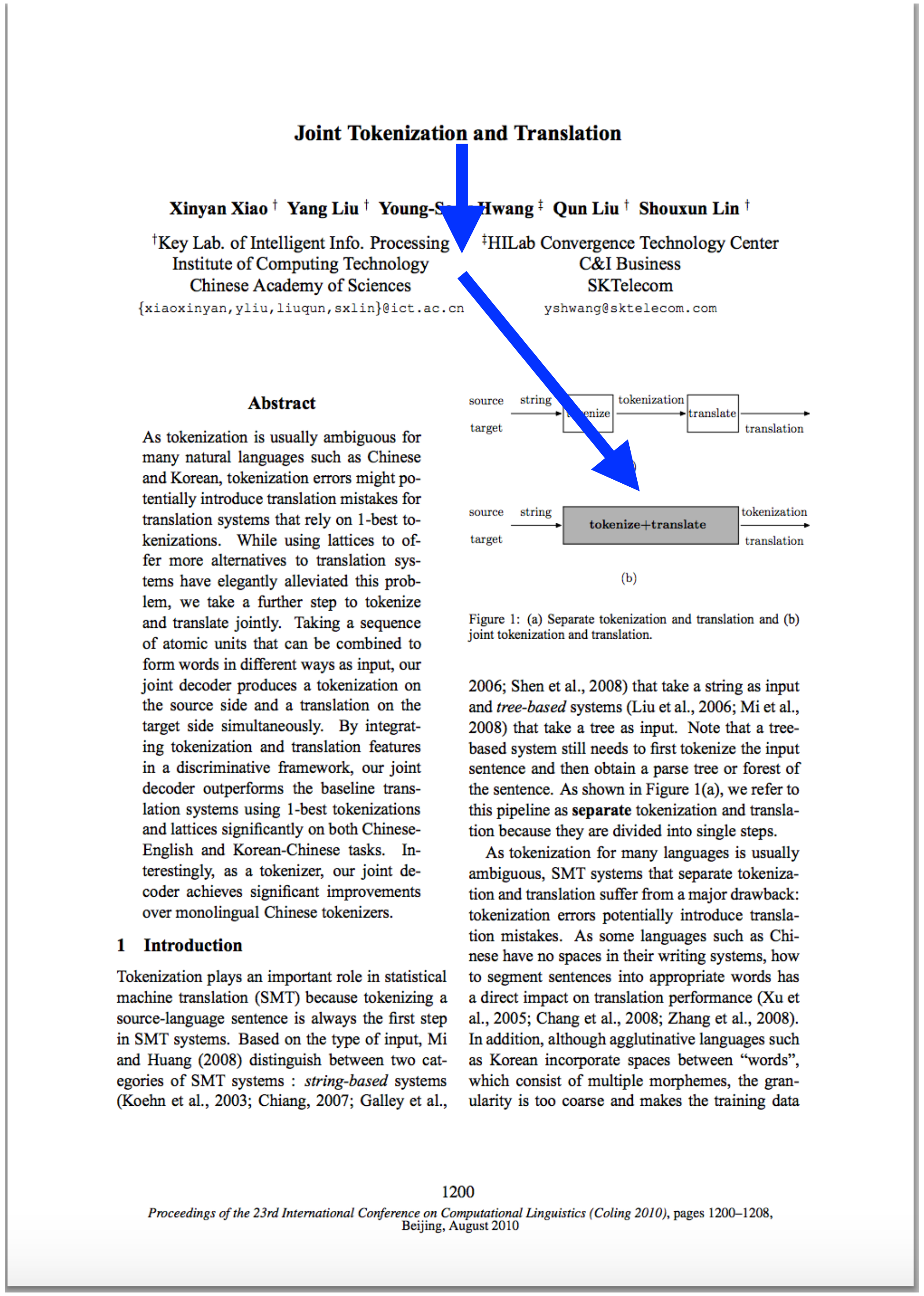
In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics



Joint Tokenization and Translation

Xinyan Xiao [†] **Yang Liu** [†] **Young-Sook Hwang** [‡] **Qun Liu** [†] **Shouxun Lin** [†]

[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn

[‡]HILab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a sequence of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

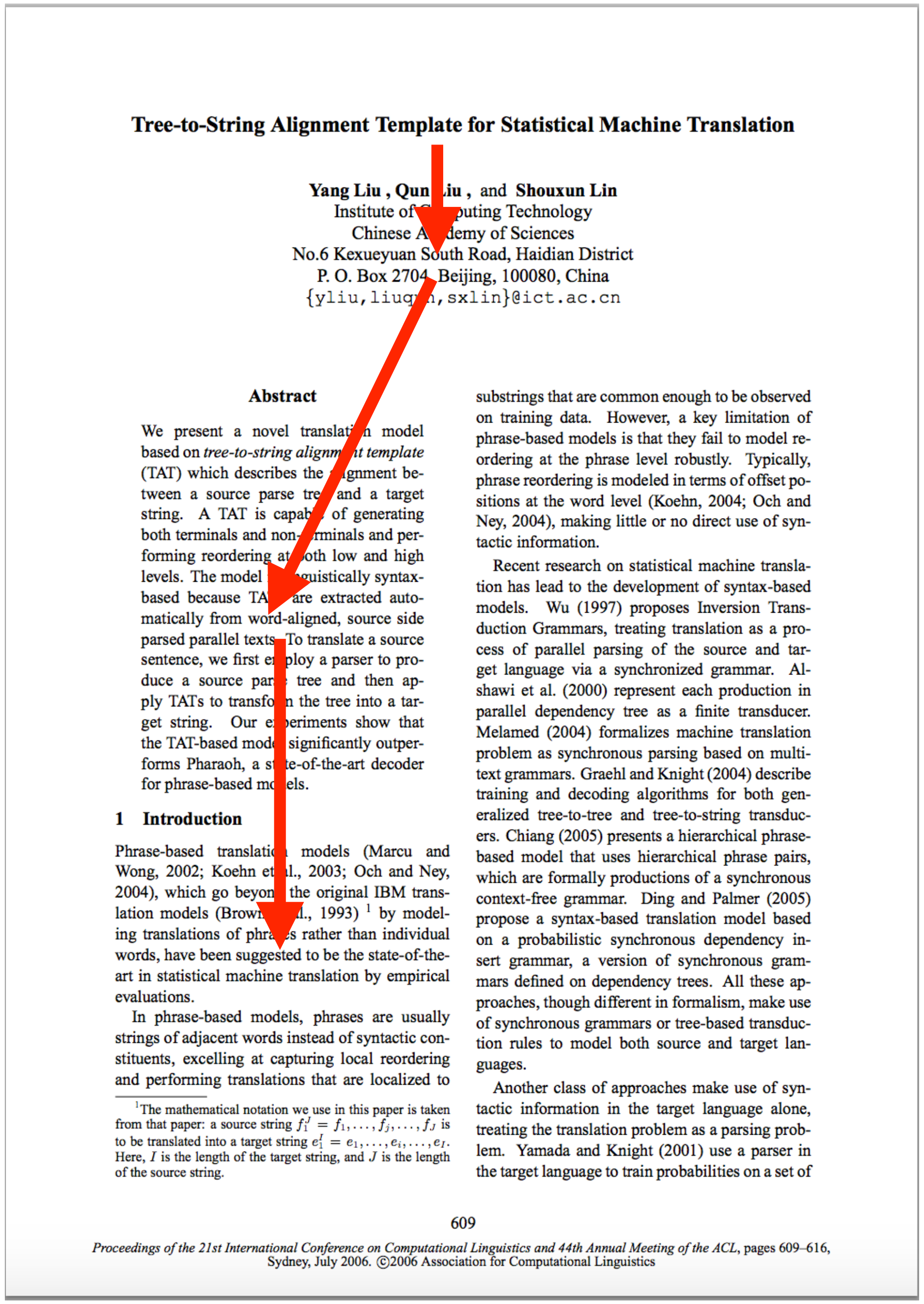
Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208,
Beijing, August 2010

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

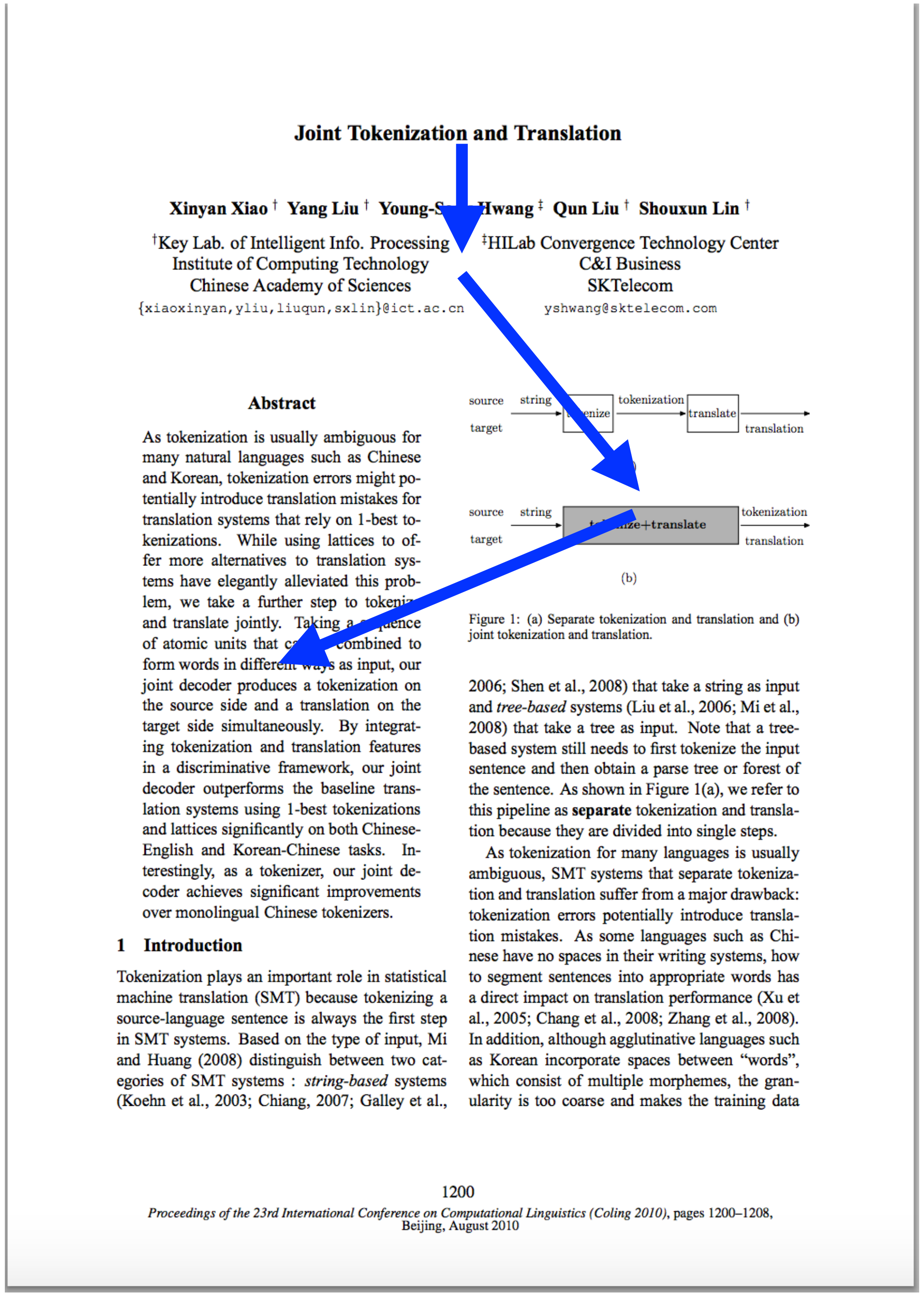
Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

609

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics



Joint Tokenization and Translation

Xinyan Xiao † **Yang Liu** † **Young-Sook Hwang** ‡ **Qun Liu** † **Shouxun Lin** †

†Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn

‡HILab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a sequence of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

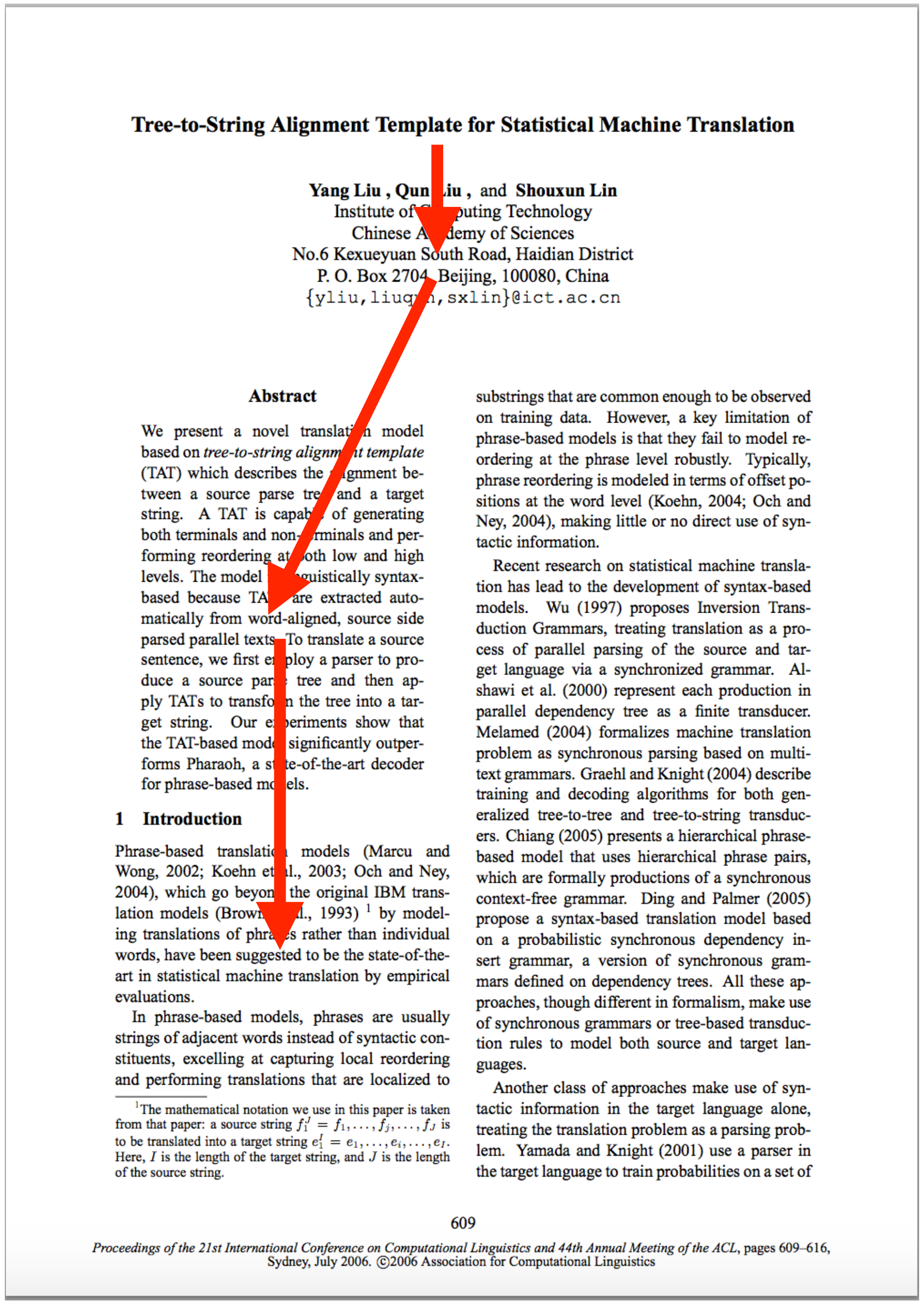
Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

1200

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208,
Beijing, August 2010

信息流的变化



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

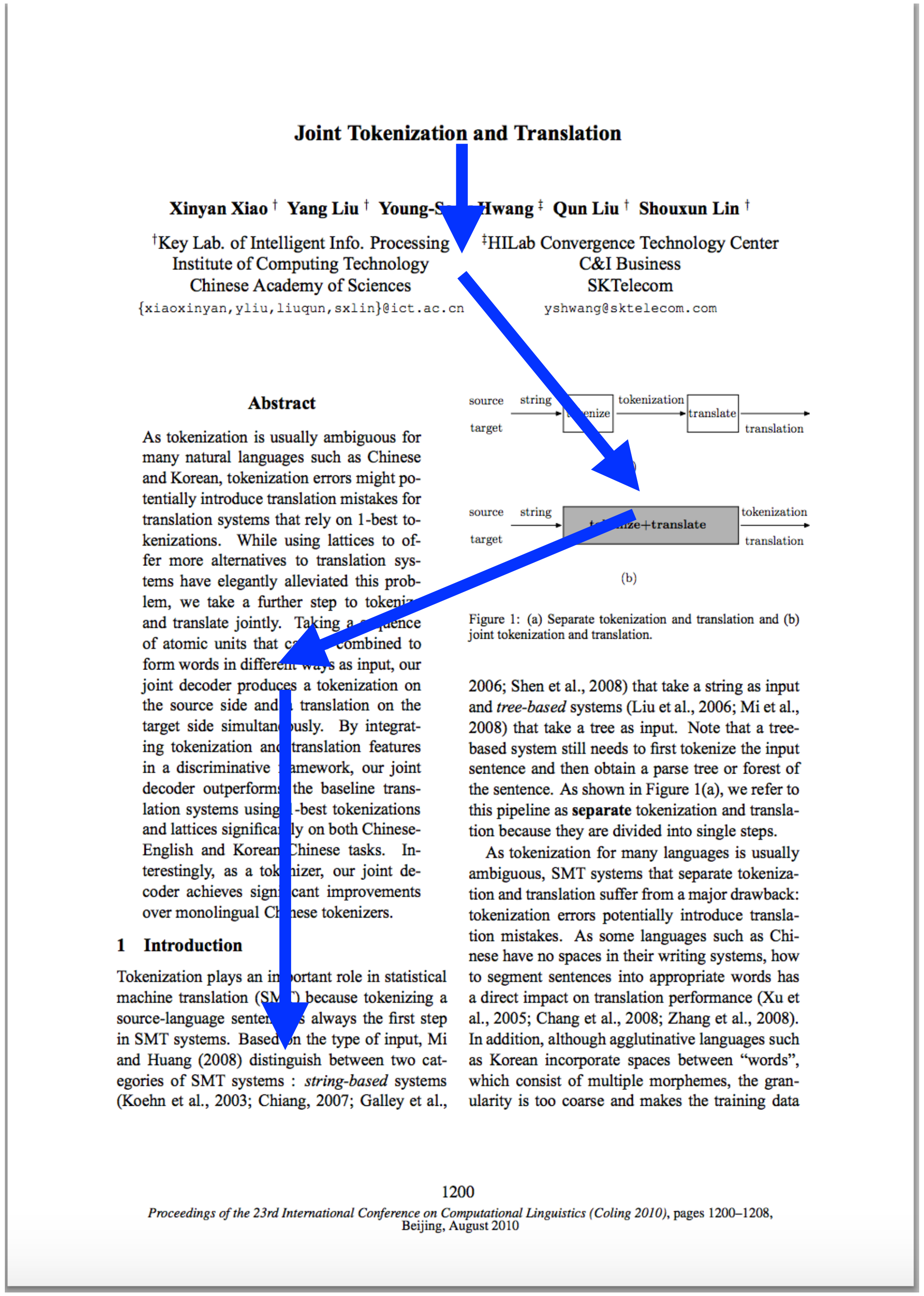
In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multitext grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616,
Sydney, July 2006. ©2006 Association for Computational Linguistics



Joint Tokenization and Translation

Xinyan Xiao [†] **Yang Liu** [†] **Young-Sook Hwang** [‡] **Qun Liu** [†] **Shouxun Lin** [†]
[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn
[‡]HILab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a sequence of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

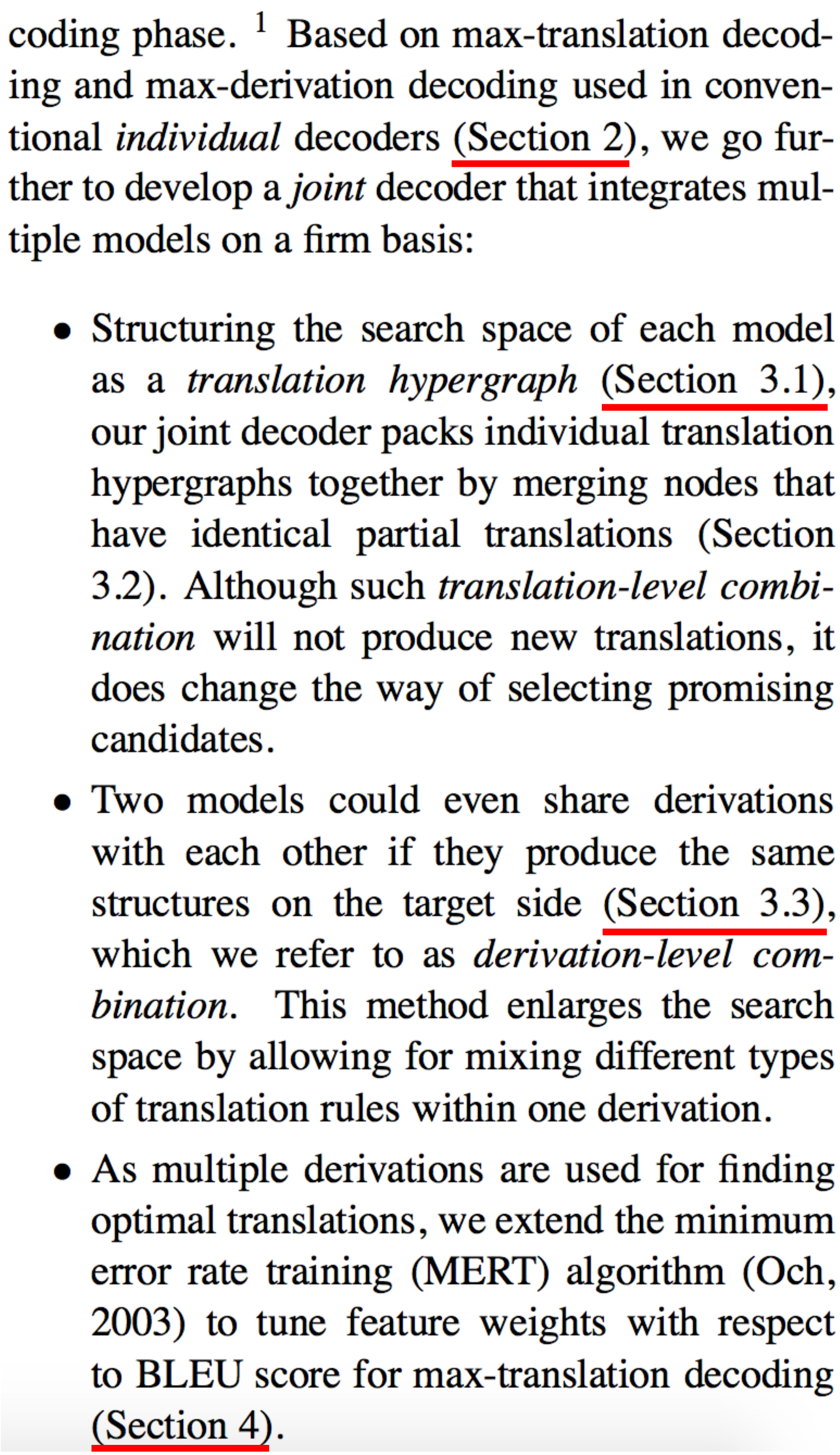
As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208,
Beijing, August 2010

图和表的重要性

- 图和表是论文的骨架，争取**让读者按照顺序看图和表就能理解论文的主要思想**，不用通过看正文才能懂
  - 一般第一遍看，都会看图、找例子
  - 然后翻到后面找主要结果
  - 再从头看正文
- 把论文的元素放在最应该被放在的地方，符合读者的认知惯性，降低理解难度

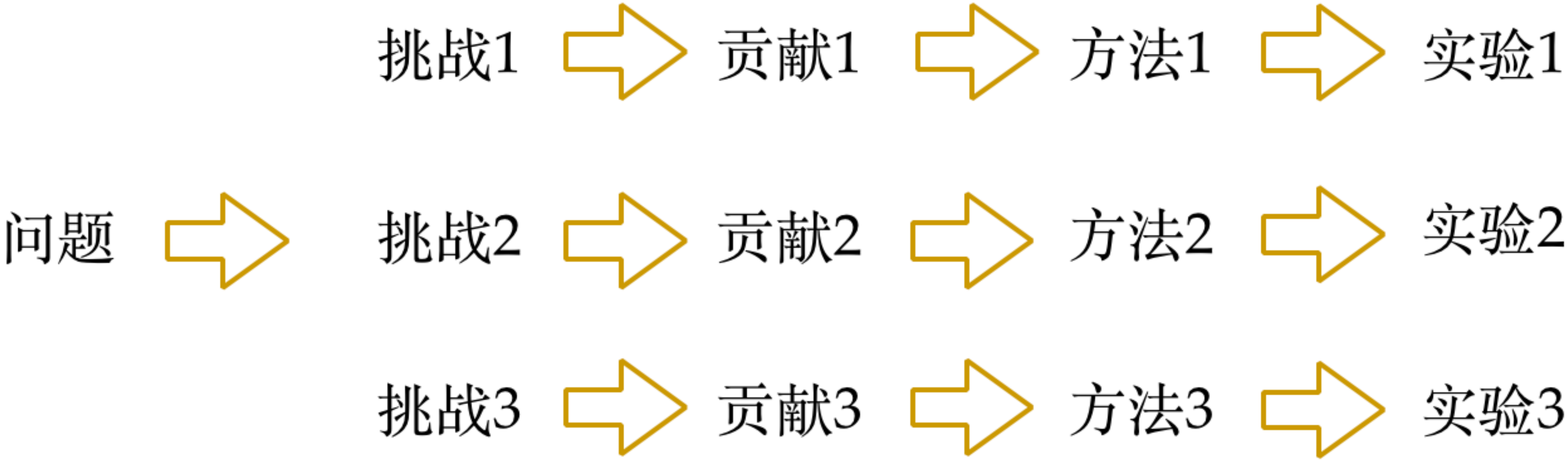
直接列出自己的贡献



coding phase. [1] Based on max-translation decoding and max-derivation decoding used in conventional *individual* decoders (Section 2), we go further to develop a *joint* decoder that integrates multiple models on a firm basis:

- Structuring the search space of each model as a *translation hypergraph* (Section 3.1), our joint decoder packs individual translation hypergraphs together by merging nodes that have identical partial translations (Section 3.2). Although such *translation-level combination* will not produce new translations, it does change the way of selecting promising candidates.
- Two models could even share derivations with each other if they produce the same structures on the target side (Section 3.3), which we refer to as *derivation-level combination*. This method enlarges the search space by allowing for mixing different types of translation rules within one derivation.
- As multiple derivations are used for finding optimal translations, we extend the minimum error rate training (MERT) algorithm (Och, 2003) to tune feature weights with respect to BLEU score for max-translation decoding (Section 4).

全局连贯性



问题

挑战1 贡献1 方法1 实验1

挑战2 贡献2 方法2 实验2

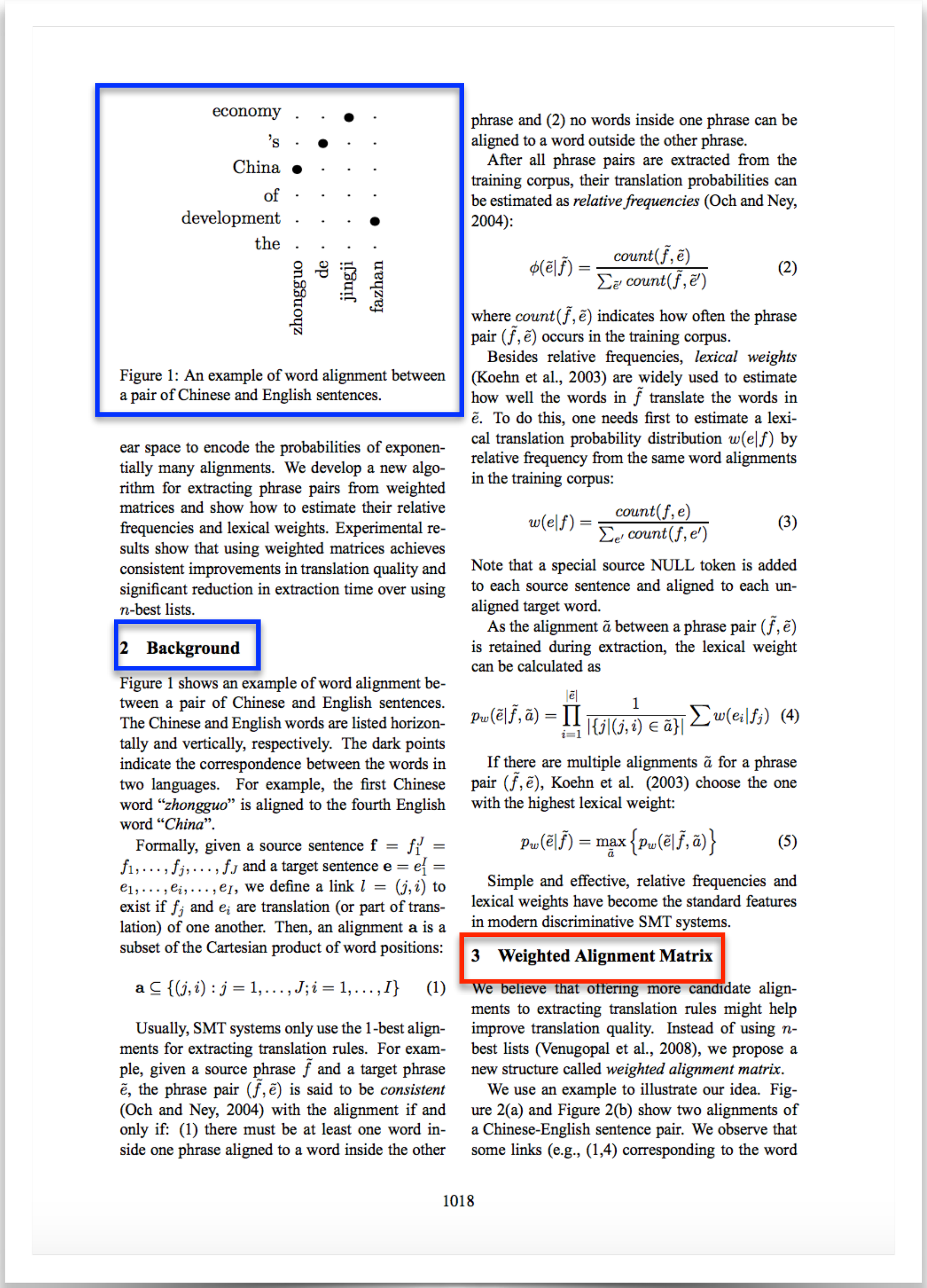
挑战3 贡献3 方法3 实验3

方法的写作技巧

如何描述自己的方法

- 不要一上来就描述你的工作，可以先介绍背景知识（往往就是baseline）
  - 有利于降低初学者或其他领域学者的理解难度
  - 有利于对introduction中的论证做更详细的解释
  - 有利于对比baseline和你的方法

例子



	zhongguo	de	jingji	fazhan
economy	·	·	●	·
’s	·	●	·	·
China	●	·	·	·
of	·	·	·	·
development	·	·	·	●
the	·	·	·	·

Figure 1: An example of word alignment between a pair of Chinese and English sentences.

ear space to encode the probabilities of exponentially many alignments. We develop a new algorithm for extracting phrase pairs from weighted matrices and show how to estimate their relative frequencies and lexical weights. Experimental results show that using weighted matrices achieves consistent improvements in translation quality and significant reduction in extraction time over using n-best lists.

2 Background

Figure 1 shows an example of word alignment between a pair of Chinese and English sentences. The Chinese and English words are listed horizontally and vertically, respectively. The dark points indicate the correspondence between the words in two languages. For example, the first Chinese word "*zhongguo*" is aligned to the fourth English word "*China*".

Formally, given a source sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$ and a target sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, e_I$, we define a link $l = (j, i)$ to exist if f_j and e_i are translation (or part of translation) of one another. Then, an alignment $\mathbf{a}$ is a subset of the Cartesian product of word positions:

$$\mathbf{a} \subseteq \{(j,i) : j = 1, \ldots, J; i = 1, \ldots, I\} \quad (1)$$

Usually, SMT systems only use the 1-best alignments for extracting translation rules. For example, given a source phrase $\tilde{f}$ and a target phrase $\tilde{e}$, the phrase pair $(\tilde{f}, \tilde{e})$ is said to be *consistent* (Och and Ney, 2004) with the alignment if and only if: (1) there must be at least one word inside one phrase aligned to a word inside the other phrase and (2) no words inside one phrase can be aligned to a word outside the other phrase.

After all phrase pairs are extracted from the training corpus, their translation probabilities can be estimated as *relative frequencies* (Och and Ney, 2004):

$$\phi(\tilde{e}|\tilde{f}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{e}'} count(\tilde{f}, \tilde{e}')} \quad (2)$$

where $count(\tilde{f}, \tilde{e})$ indicates how often the phrase pair $(\tilde{f}, \tilde{e})$ occurs in the training corpus.

Besides relative frequencies, *lexical weights* (Koehn et al., 2003) are widely used to estimate how well the words in $\tilde{f}$ translate the words in $\tilde{e}$. To do this, one needs first to estimate a lexical translation probability distribution $w(e|f)$ by relative frequency from the same word alignments in the training corpus:

$$w(e|f) = \frac{count(f, e)}{\sum_{e'} count(f, e')} \quad (3)$$

Note that a special source NULL token is added to each source sentence and aligned to each unaligned target word.

As the alignment $\tilde{a}$ between a phrase pair $(\tilde{f}, \tilde{e})$ is retained during extraction, the lexical weight can be calculated as

$$p_w(\tilde{e}|\tilde{f}, \tilde{a}) = \prod_{i=1}^{|\tilde{e}|} \frac{1}{|\{j|(j,i) \in \tilde{a}\}|} \sum w(e_i|f_j) \quad (4)$$

If there are multiple alignments $\tilde{a}$ for a phrase pair $(\tilde{f}, \tilde{e})$, Koehn et al. (2003) choose the one with the highest lexical weight:

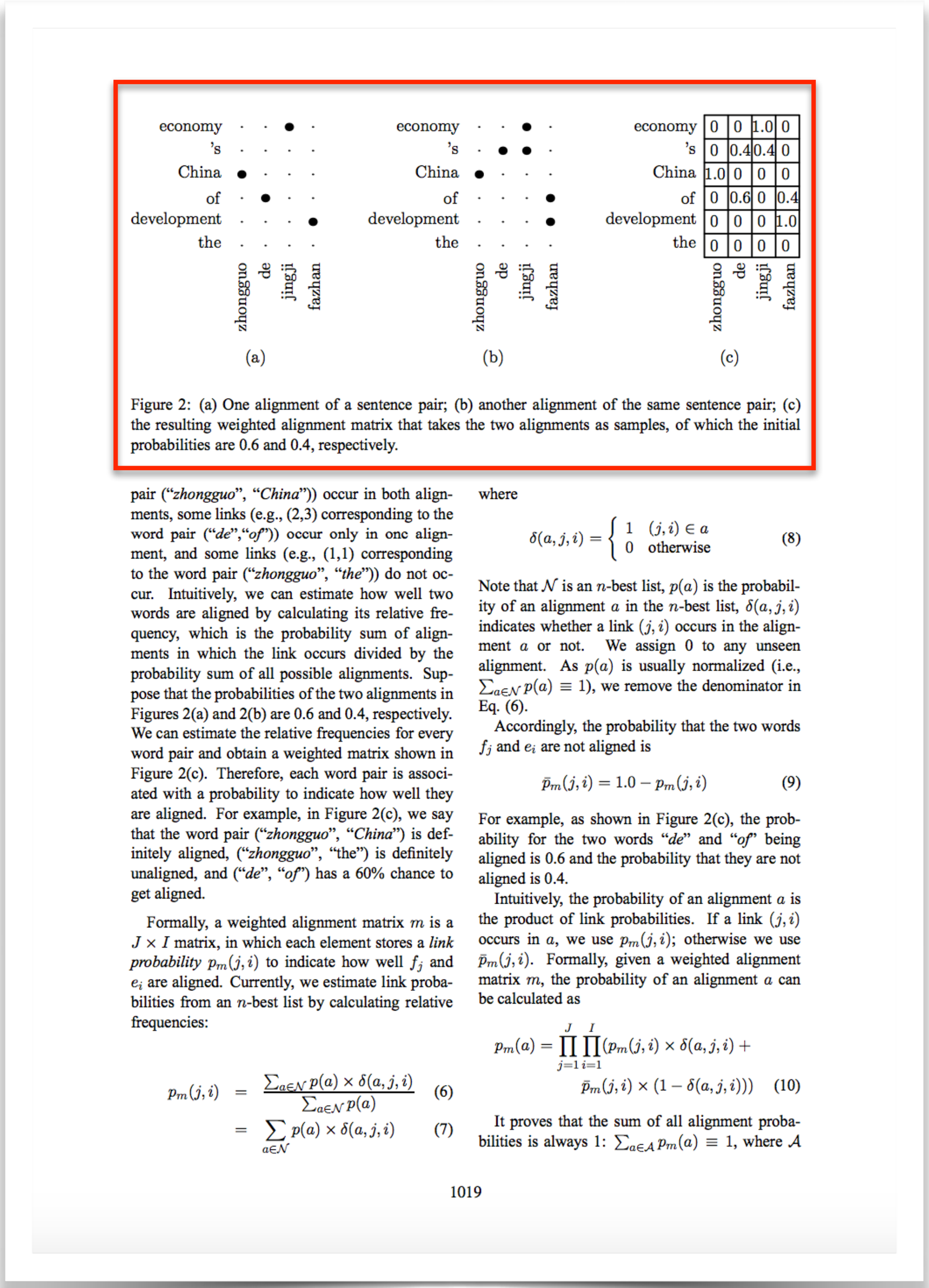
$$p_w(\tilde{e}|\tilde{f}) = \max_{\tilde{a}} \left\{ p_w(\tilde{e}|\tilde{f}, \tilde{a}) \right\} \quad (5)$$

Simple and effective, relative frequencies and lexical weights have become the standard features in modern discriminative SMT systems.

3 Weighted Alignment Matrix

We believe that offering more candidate alignments to extracting translation rules might help improve translation quality. Instead of using n-best lists (Venugopal et al., 2008), we propose a new structure called *weighted alignment matrix*.

We use an example to illustrate our idea. Figure 2(a) and Figure 2(b) show two alignments of a Chinese-English sentence pair. We observe that some links (e.g., (1,4) corresponding to the word



(a)

	zhongguo	de	jingji	fazhan
economy	·	·	●	·
’s	·	·	·	·
China	●	·	·	·
of	·	●	·	·
development	·	·	·	●
the	·	·	·	·

(b)

	zhongguo	de	jingji	fazhan
economy	·	·	●	·
’s	·	●	●	·
China	●	·	·	·
of	·	·	·	●
development	·	·	·	●
the	·	·	·	·

(c)

	zhongguo	de	jingji	fazhan
economy	0	0	1.0	0
’s	0	0.4	0.4	0
China	1.0	0	0	0
of	0	0.6	0	0.4
development	0	0	0	1.0
the	0	0	0	0

Figure 2: (a) One alignment of a sentence pair; (b) another alignment of the same sentence pair; (c) the resulting weighted alignment matrix that takes the two alignments as samples, of which the initial probabilities are 0.6 and 0.4, respectively.

pair ("*zhongguo*", "*China*")) occur in both alignments, some links (e.g., (2,3) corresponding to the word pair ("*de*","*of*")) occur only in onc alignment, and some links (e.g., (1,1) corresponding to the word pair ("*zhongguo*", "*the*")) do not occur. Intuitively, we can estimate how well two words are aligned by calculating its relative frequency, which is the probability sum of alignments in which the link occurs divided by the probability sum of all possible alignments. Suppose that the probabilities of the two alignments in Figures 2(a) and 2(b) are 0.6 and 0.4, respectively. We can estimate the relative frequencies for every word pair and obtain a weighted matrix shown in Figure 2(c). Therefore, each word pair is associated with a probability to indicate how well they are aligned. For example, in Figure 2(c), we say that the word pair ("*zhongguo*", "*China*") is definitely aligned, ("*zhongguo*", "the") is definitely unaligned, and ("*de*", "*of*") has a 60% chance to get aligned.

Formally, a weighted alignment matrix m is a $J \times I$ matrix, in which each element stores a *link probability* $p_m(j,i)$ to indicate how well f_j and e_i are aligned. Currently, we estimate link probabilities from an n-best list by calculating relative frequencies:

$$\begin{aligned} p_m(j,i) &= \frac{\sum_{a \in \mathcal{N}} p(a) \times \delta(a,j,i)}{\sum_{a \in \mathcal{N}} p(a)} && (6) \\ &= \sum_{a \in \mathcal{N}} p(a) \times \delta(a,j,i) && (7) \end{aligned}$$

where

$$\delta(a,j,i) = \begin{cases} 1 & (j,i) \in a \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Note that $\mathcal{N}$ is an n-best list, $p(a)$ is the probability of an alignment a in the n-best list, $\delta(a,j,i)$ indicates whether a link (j,i) occurs in the alignment a or not. We assign 0 to any unseen alignment. As $p(a)$ is usually normalized (i.e., $\sum_{a \in \mathcal{N}} p(a) \equiv 1$), we remove the denominator in Eq. (6).

Accordingly, the probability that the two words f_j and e_i are not aligned is

$$\bar{p}_m(j,i) = 1.0 - p_m(j,i) \quad (9)$$

For example, as shown in Figure 2(c), the probability for the two words "*de*" and "*of*" being aligned is 0.6 and the probability that they are not aligned is 0.4.

Intuitively, the probability of an alignment a is the product of link probabilities. If a link (j,i) occurs in a, we use $p_m(j,i)$; otherwise we use $\bar{p}_m(j,i)$. Formally, given a weighted alignment matrix m, the probability of an alignment a can be calculated as

$$\begin{aligned} p_m(a) = \prod_{j=1}^{J} \prod_{i=1}^{I} (p_m(j,i) \times \delta(a,j,i) + \\ \bar{p}_m(j,i) \times (1 - \delta(a,j,i))) \quad (10) \end{aligned}$$

It proves that the sum of all alignment probabilities is always 1: $\sum_{a \in \mathcal{A}} p_m(a) \equiv 1$, where $\mathcal{A}$

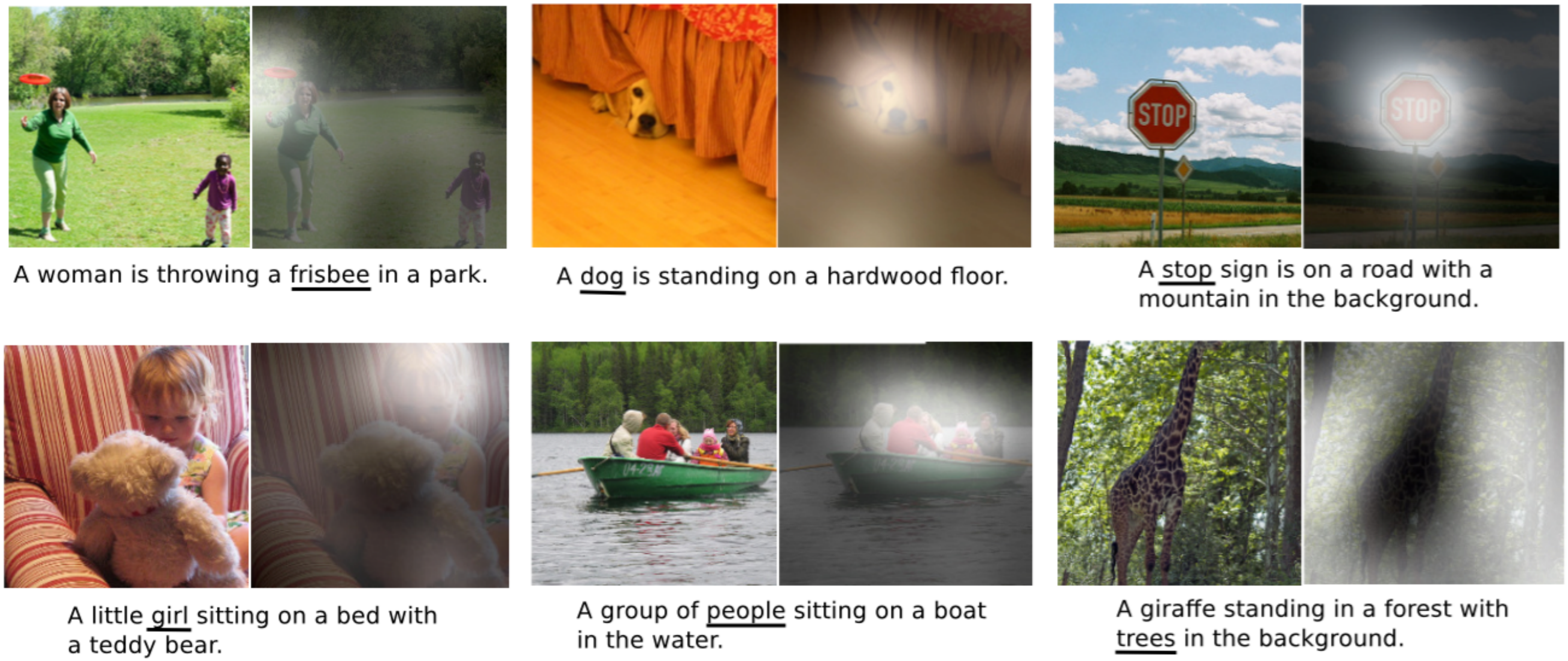
Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. **Weighted Alignment Matrices for Statistical Machine Translation**. In *EMNLP 2009*.

Running Example是利器

- 英语不好说不清楚？用例子！
- 全篇统一使用一个running example，用来阐释你的方法（甚至是baseline）
- 围绕着running example，展开描述你的工作
- 审稿人能从running example中更舒服地了解你的工作，读正文会花掉他/她更多时间
- 看完running example，审稿人便能知道核心思想

一图胜千言

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Joshua Bengio. **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention**. In *ICML 2015*.

方法描述的逻辑顺序

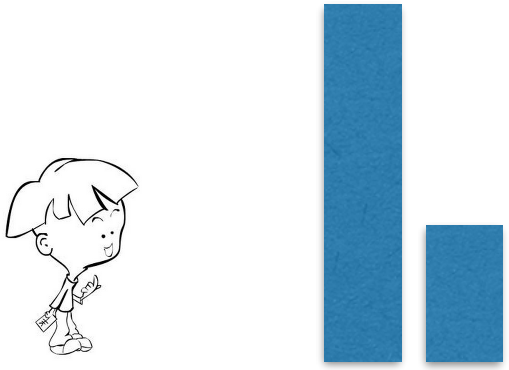
- 错误的顺序
    - 形式化描述
    - 解释数学符号的意义

方法描述的逻辑顺序

- 错误的顺序
  - 形式化描述
  - 解释数学符号的意义
- 正确的顺序
  - 首先给出running example
  - 然后利用running example，用通俗语言描述你的想法
  - 最后才是形式化描述

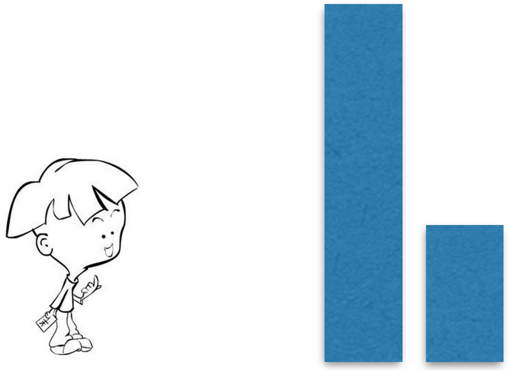
方法描述的逻辑顺序

- 错误的顺序
  - 形式化描述
  - 解释数学符号的意义
- 正确的顺序
  - 首先给出running example
  - 然后利用running example，用通俗语言描述你的想法
  - 最后才是形式化描述

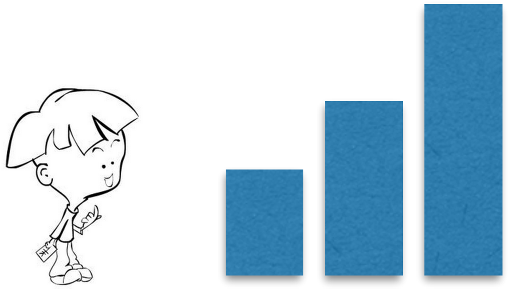


方法描述的逻辑顺序

- 错误的顺序
  - 形式化描述
  - 解释数学符号的意义

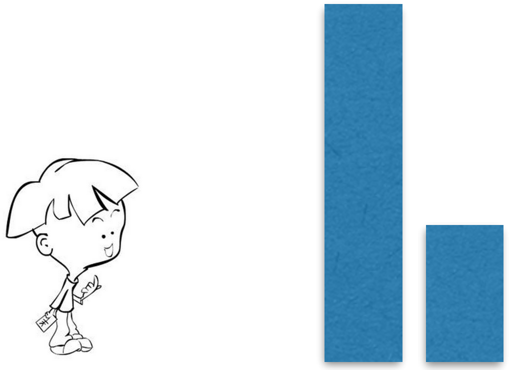


- 正确的顺序
  - 首先给出running example
  - 然后利用running example，用通俗语言描述你的想法
  - 最后才是形式化描述

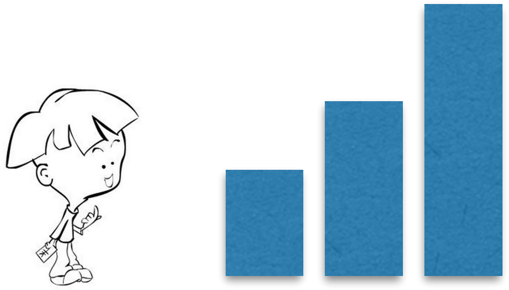


方法描述的逻辑顺序

- 错误的顺序
  - 形式化描述
  - 解释数学符号的意义

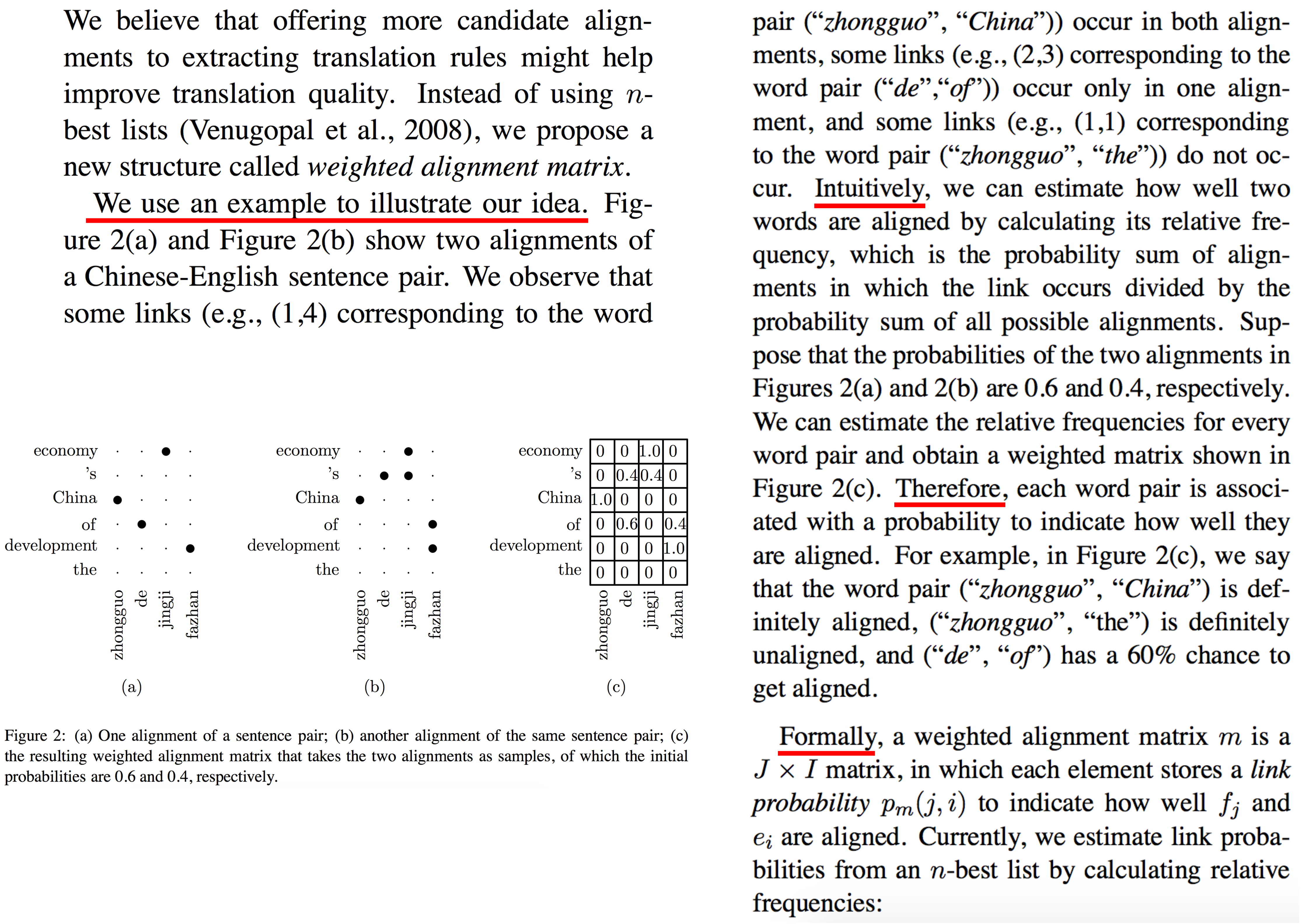


- 正确的顺序
  - 首先给出running example
  - 然后利用running example，用通俗语言描述你的想法
  - 最后才是形式化描述



每个公式都有语言学意义，都来自你的直觉和想法，直接告诉审稿人，不要让他/她去揣摩

例子



We believe that offering more candidate alignments to extracting translation rules might help improve translation quality. Instead of using n-best lists (Venugopal et al., 2008), we propose a new structure called *weighted alignment matrix*.

We use an example to illustrate our idea. Figure 2(a) and Figure 2(b) show two alignments of a Chinese-English sentence pair. We observe that some links (e.g., (1,4) corresponding to the word pair ("*zhongguo*", "*China*")) occur in both alignments, some links (e.g., (2,3) corresponding to the word pair ("*de*","*of*")) occur only in one alignment, and some links (e.g., (1,1) corresponding to the word pair ("*zhongguo*", "*the*")) do not occur. Intuitively, we can estimate how well two words are aligned by calculating its relative frequency, which is the probability sum of alignments in which the link occurs divided by the probability sum of all possible alignments. Suppose that the probabilities of the two alignments in Figures 2(a) and 2(b) are 0.6 and 0.4, respectively. We can estimate the relative frequencies for every word pair and obtain a weighted matrix shown in Figure 2(c). Therefore, each word pair is associated with a probability to indicate how well they are aligned. For example, in Figure 2(c), we say that the word pair ("*zhongguo*", "*China*") is definitely aligned, ("*zhongguo*", "the") is definitely unaligned, and ("*de*", "*of*") has a 60% chance to get aligned.

economy	·	·	●	·
's	·	·	·	·
China	●	·	·	·
of	·	●	·	·
development	·	·	·	●
the	·	·	·	·
	zhongguo	de	jingji	fazhan

(a)

economy	·	·	●	·
's	·	●	●	·
China	●	·	·	·
of	·	·	·	●
development	·	·	·	●
the	·	·	·	·
	zhongguo	de	jingji	fazhan

(b)

economy	0	0	1.0	0
's	0	0.4	0.4	0
China	1.0	0	0	0
of	0	0.6	0	0.4
development	0	0	0	1.0
the	0	0	0	0
	zhongguo	de	jingji	fazhan

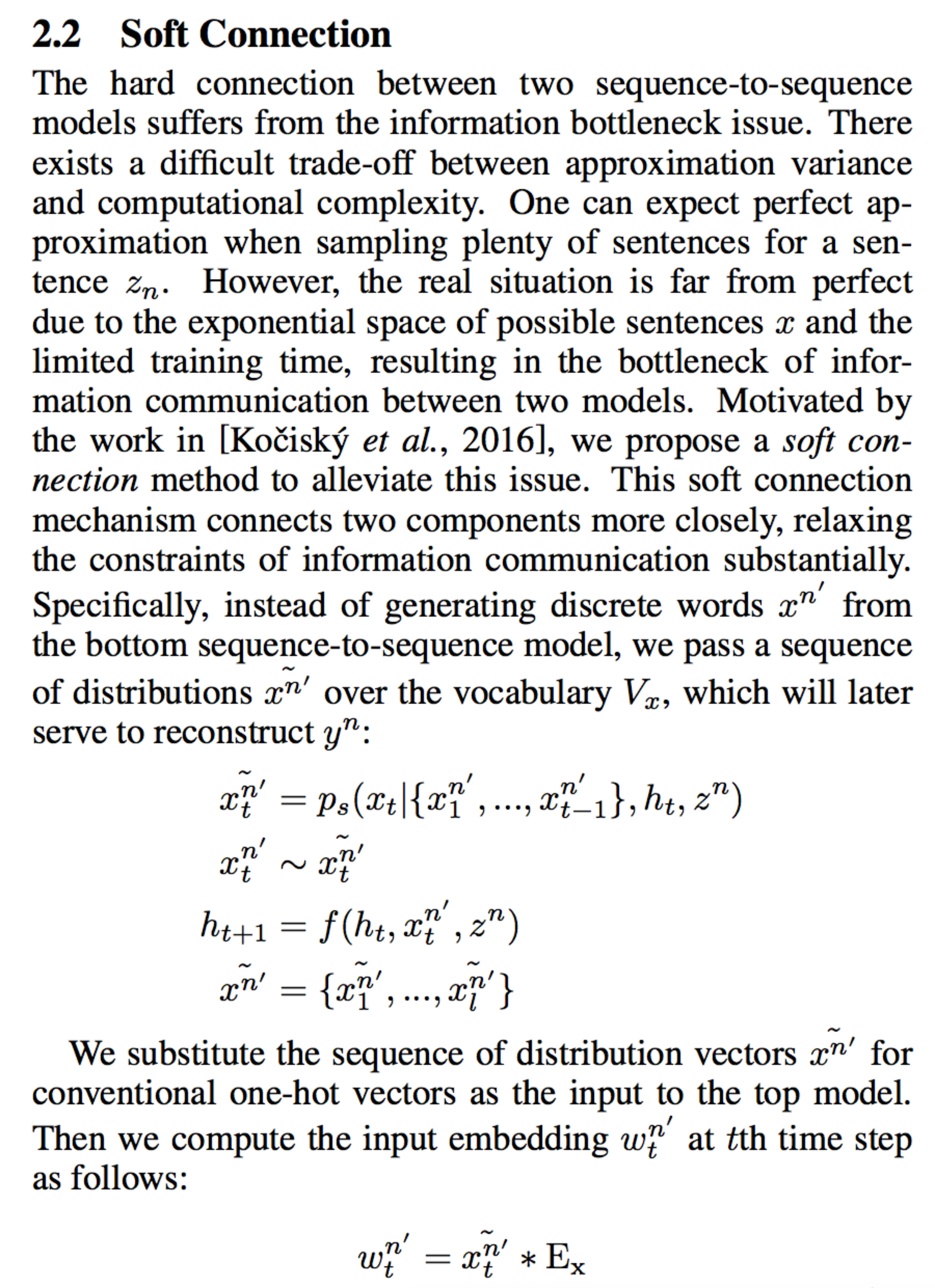
(c)

Figure 2: (a) One alignment of a sentence pair; (b) another alignment of the same sentence pair; (c) the resulting weighted alignment matrix that takes the two alignments as samples, of which the initial probabilities are 0.6 and 0.4, respectively.

Formally, a weighted alignment matrix m is a $J \times I$ matrix, in which each element stores a *link probability* $p_m(j, i)$ to indicate how well f_j and e_i are aligned. Currently, we estimate link probabilities from an n-best list by calculating relative frequencies:

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. **Weighted Alignment Matrices for Statistical Machine Translation**. In EMNLP *2009*.

描述的准确性与形式化能力



2.2 Soft Connection

The hard connection between two sequence-to-sequence models suffers from the information bottleneck issue. There exists a difficult trade-off between approximation variance and computational complexity. One can expect perfect approximation when sampling plenty of sentences for a sentence z_n. However, the real situation is far from perfect due to the exponential space of possible sentences x and the limited training time, resulting in the bottleneck of information communication between two models. Motivated by the work in [Kočiský *et al.*, 2016], we propose a *soft connection* method to alleviate this issue. This soft connection mechanism connects two components more closely, relaxing the constraints of information communication substantially. Specifically, instead of generating discrete words $x^{n'}$ from the bottom sequence-to-sequence model, we pass a sequence of distributions $\tilde{x}^{n'}$ over the vocabulary V_x, which will later serve to reconstruct y^n:

$$\tilde{x}_t^{n'} = p_s(x_t|\{x_1^{n'}, ..., x_{t-1}^{n'}\}, h_t, z^n)$$

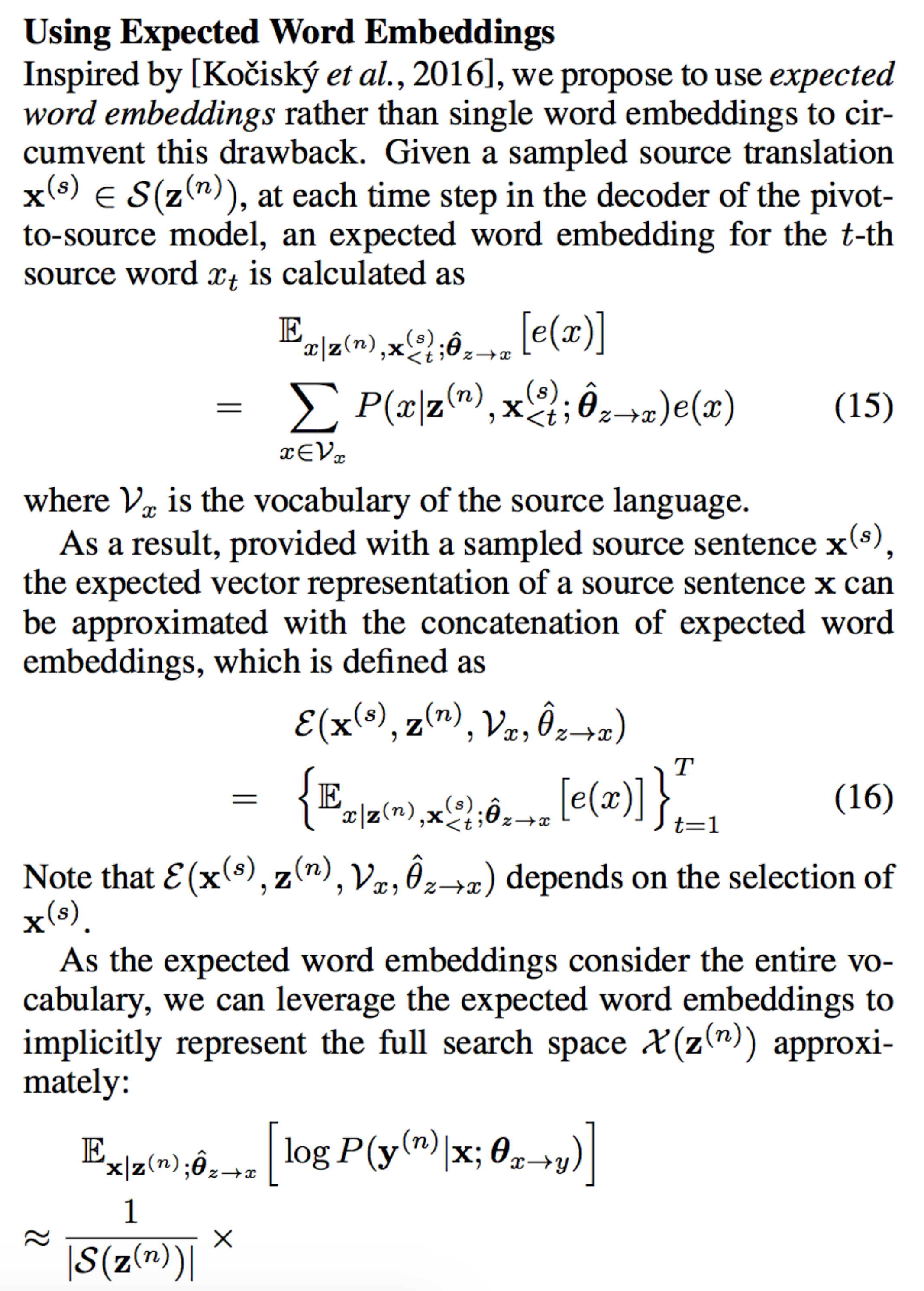
$$x_t^{n'} \sim \tilde{x}_t^{n'}$$

$$h_{t+1} = f(h_t, x_t^{n'}, z^n)$$

$$\tilde{x}^{n'} = \{\tilde{x}_1^{n'}, ..., \tilde{x}_l^{n'}\}$$

We substitute the sequence of distribution vectors $\tilde{x}^{n'}$ for conventional one-hot vectors as the input to the top model. Then we compute the input embedding $w_t^{n'}$ at tth time step as follows:

$$w_t^{n'} = \tilde{x}_t^{n'} * \mathrm{E_x}$$



Using Expected Word Embeddings

Inspired by [Kočiský *et al.*, 2016], we propose to use *expected word embeddings* rather than single word embeddings to circumvent this drawback. Given a sampled source translation $\mathbf{x}^{(s)} \in \mathcal{S}(\mathbf{z}^{(n)})$, at each time step in the decoder of the pivot-to-source model, an expected word embedding for the t-th source word x_t is calculated as

$$\begin{aligned} & \mathbb{E}_{x|\mathbf{z}^{(n)},\mathbf{x}^{(s)}_{<t};\hat{\boldsymbol{\theta}}_{z\rightarrow x}}\big[e(x)\big] \\ = & \sum_{x \in \mathcal{V}_x} P(x|\mathbf{z}^{(n)}, \mathbf{x}^{(s)}_{<t}; \hat{\boldsymbol{\theta}}_{z\rightarrow x})e(x) \end{aligned} \qquad (15)$$

where $\mathcal{V}_x$ is the vocabulary of the source language.

As a result, provided with a sampled source sentence $\mathbf{x}^{(s)}$, the expected vector representation of a source sentence $\mathbf{x}$ can be approximated with the concatenation of expected word embeddings, which is defined as

$$\begin{aligned} & \mathcal{E}(\mathbf{x}^{(s)}, \mathbf{z}^{(n)}, \mathcal{V}_x, \hat{\theta}_{z\rightarrow x}) \\ = & \Big\{\mathbb{E}_{x|\mathbf{z}^{(n)},\mathbf{x}^{(s)}_{<t};\hat{\boldsymbol{\theta}}_{z\rightarrow x}}\big[e(x)\big]\Big\}_{t=1}^{T} \end{aligned} \qquad (16)$$

Note that $\mathcal{E}(\mathbf{x}^{(s)}, \mathbf{z}^{(n)}, \mathcal{V}_x, \hat{\theta}_{z\rightarrow x})$ depends on the selection of $\mathbf{x}^{(s)}$.

As the expected word embeddings consider the entire vocabulary, we can leverage the expected word embeddings to implicitly represent the full search space $\mathcal{X}(\mathbf{z}^{(n)})$ approximately:

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\rightarrow x}}\Big[\log P(\mathbf{y}^{(n)}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow y})\Big] \\ \approx & \frac{1}{|\mathcal{S}(\mathbf{z}^{(n)})|} \times \end{aligned}$$

Hao Zheng, Long Cheng, and Yang Liu. **Maximum Expected Likelihood Estimation for Neural Machine Translation**. In *IJCAI 2017*.

实验的写作技巧

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
  - 辅助实验（开发集）：参数的影响
  - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
  - 辅助实验（开发集）：参数的影响
  - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

minimum

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
    - 辅助实验（开发集）：参数的影响
    - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

minimum ⇨

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
  - 辅助实验（开发集）：参数的影响
  - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

minimum ⇨ solid

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
  - 辅助实验（开发集）：参数的影响
  - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

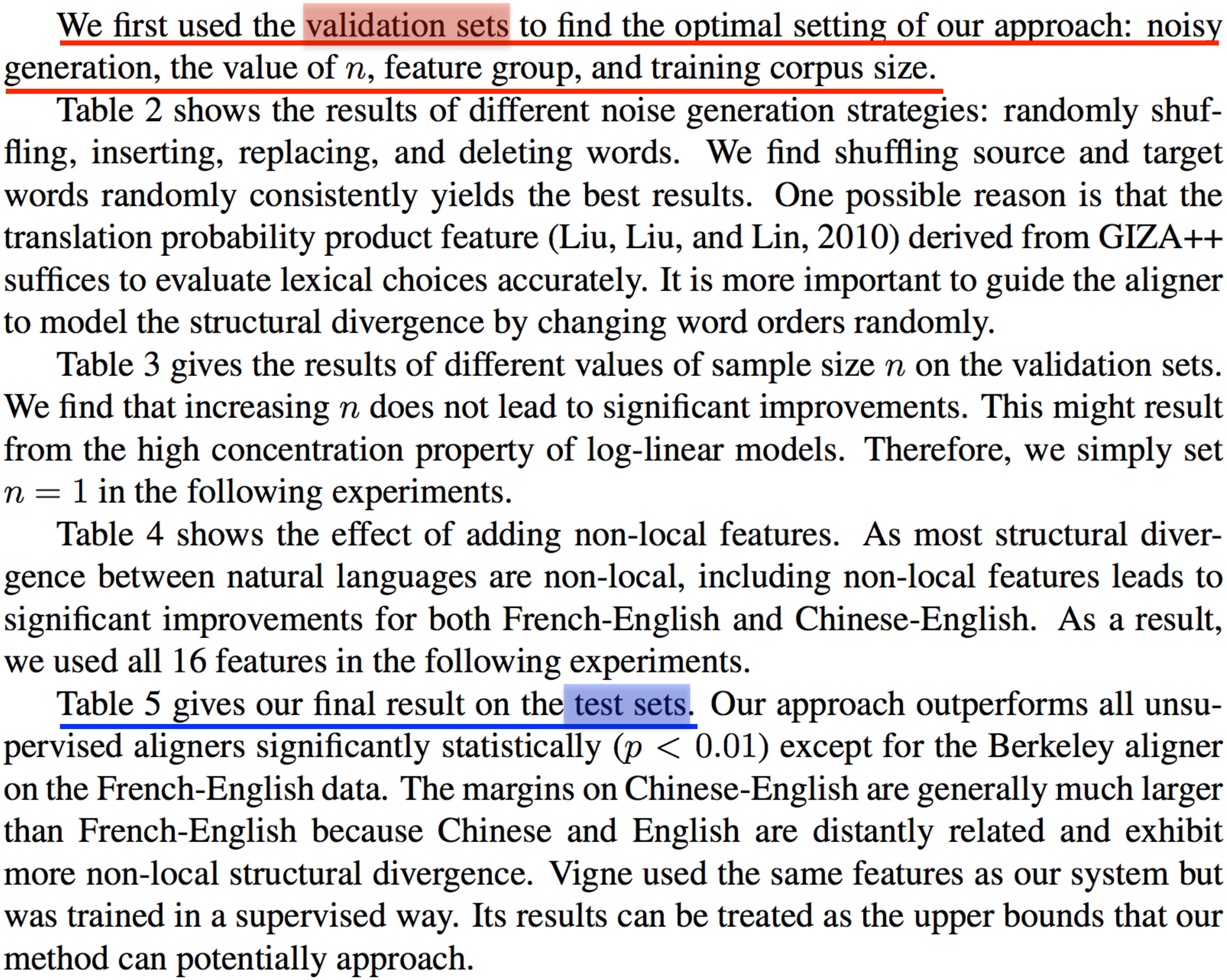
minimum ⇨ solid ⇨

实验设计

- 公认的标准数据和state-of-the-art系统
- 实验先辅后主
    - 辅助实验（开发集）：参数的影响
    - 主实验（测试集）：证明显著超过baseline
- 必须有显著性检验
- 不辞辛劳，做到极致

minimum ⇨ solid ⇨ maximum

先辅后主



We first used the validation sets to find the optimal setting of our approach: noisy generation, the value of n, feature group, and training corpus size.

Table 2 shows the results of different noise generation strategies: randomly shuffling, inserting, replacing, and deleting words. We find shuffling source and target words randomly consistently yields the best results. One possible reason is that the translation probability product feature (Liu, Liu, and Lin, 2010) derived from GIZA++ suffices to evaluate lexical choices accurately. It is more important to guide the aligner to model the structural divergence by changing word orders randomly.

Table 3 gives the results of different values of sample size n on the validation sets. We find that increasing n does not lead to significant improvements. This might result from the high concentration property of log-linear models. Therefore, we simply set $n = 1$ in the following experiments.

Table 4 shows the effect of adding non-local features. As most structural divergence between natural languages are non-local, including non-local features leads to significant improvements for both French-English and Chinese-English. As a result, we used all 16 features in the following experiments.

Table 5 gives our final result on the test sets. Our approach outperforms all unsupervised aligners significantly statistically ($p < 0.01$) except for the Berkeley aligner on the French-English data. The margins on Chinese-English are generally much larger than French-English because Chinese and English are distantly related and exhibit more non-local structural divergence. Vigne used the same features as our system but was trained in a supervised way. Its results can be treated as the upper bounds that our method can potentially approach.

Yang Liu and Massing Sun. **Contrastive Unsupervised Word Alignment with Non-Local Features**. *arXiv:1410.2082[cs.CL]*.

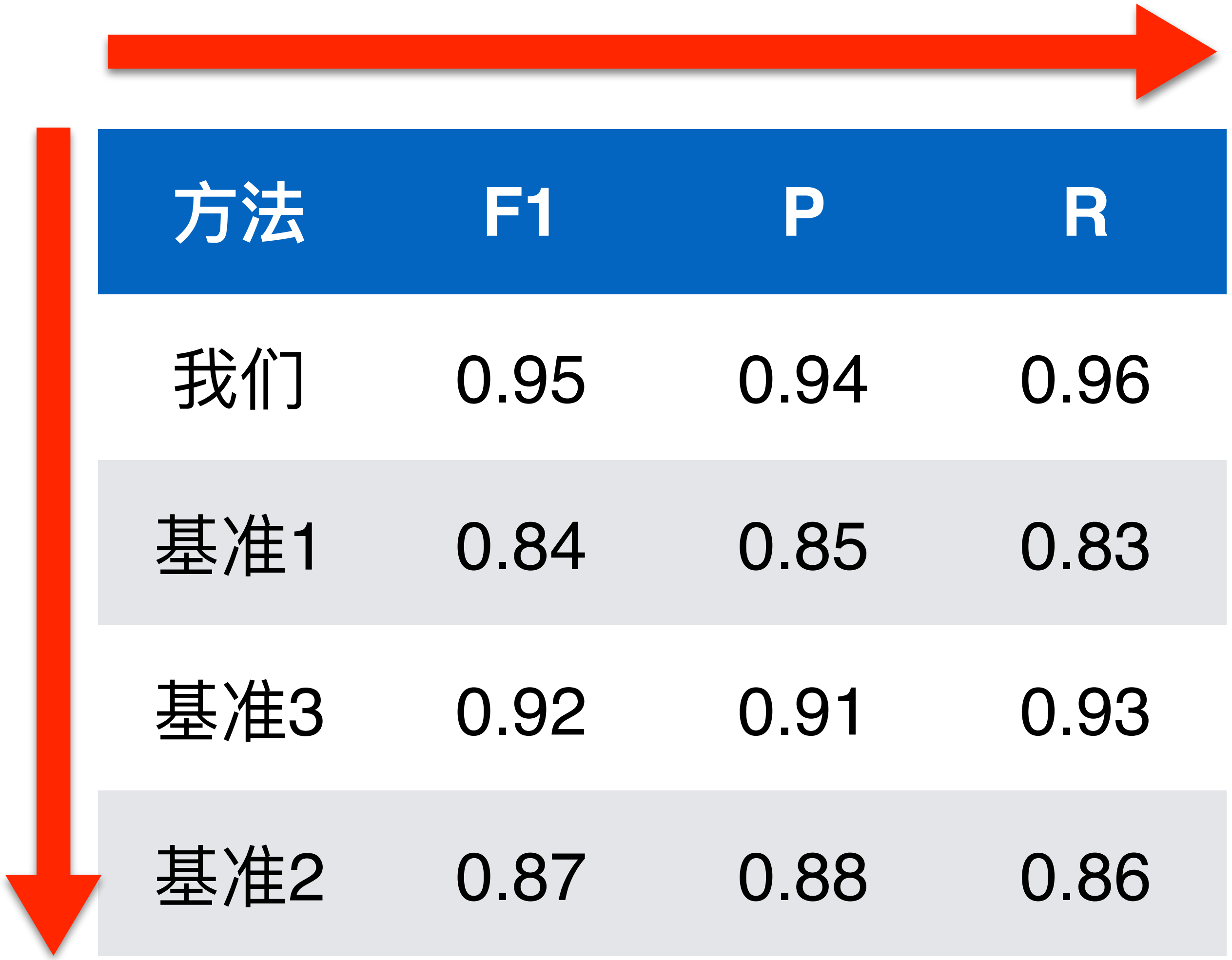
用表的误区

方法	F1	P	R
我们	0.95	0.94	0.96
基准1	0.84	0.85	0.83
基准3	0.92	0.91	0.93
基准2	0.87	0.88	0.86

用表的误区

方法	F1	P	R
我们	0.95	0.94	0.96
基准1	0.84	0.85	0.83
基准3	0.92	0.91	0.93
基准2	0.87	0.88	0.86

用表的误区



方法	F1	P	R
我们	0.95	0.94	0.96
基准1	0.84	0.85	0.83
基准3	0.92	0.91	0.93
基准2	0.87	0.88	0.86

用表的误区

方法	F1	P	R
我们	0.95	0.94	0.96
基准1	0.84	0.85	0.83
基准3	0.92	0.91	0.93
基准2	0.87	0.88	0.86

阅读顺序：从上向下，从左向右

baseline在上，我们的方法在下，最终结果在最后一列

用表的误区

方法	F1	P	R
我们	0.95	0.94	0.96
基准1	0.84	0.85	0.83
基准3	0.92	0.91	0.93
基准2	0.87	0.88	0.86

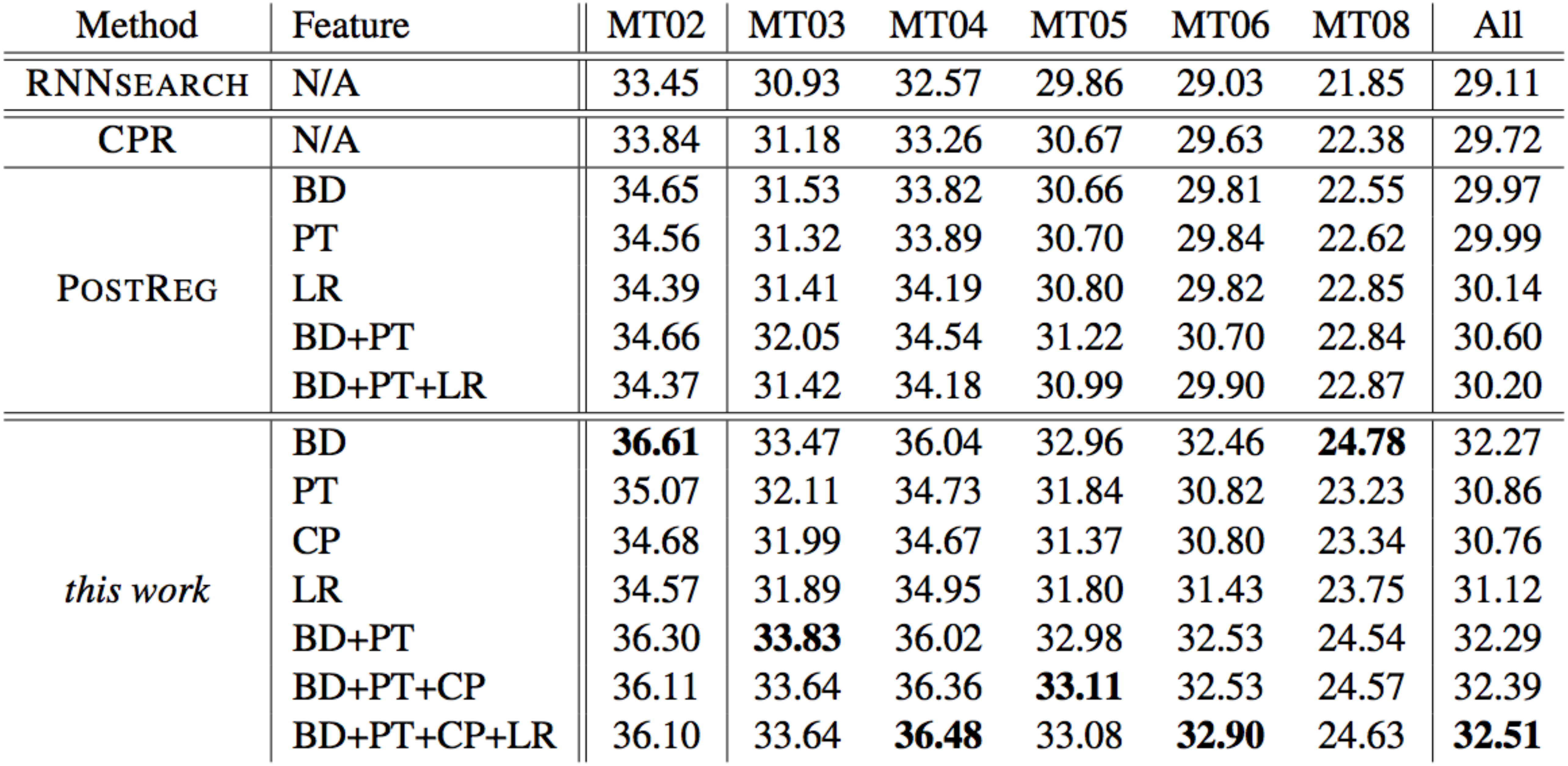
方法	P	R	F
基准1	0.85	0.83	0.84
基准2	0.87	0.88	0.86
基准3	0.92	0.91	0.93
我们	0.95	0.94	0.96

阅读顺序：从上向下，从左向右

baseline在上，我们的方法在下，最终结果在最后一列

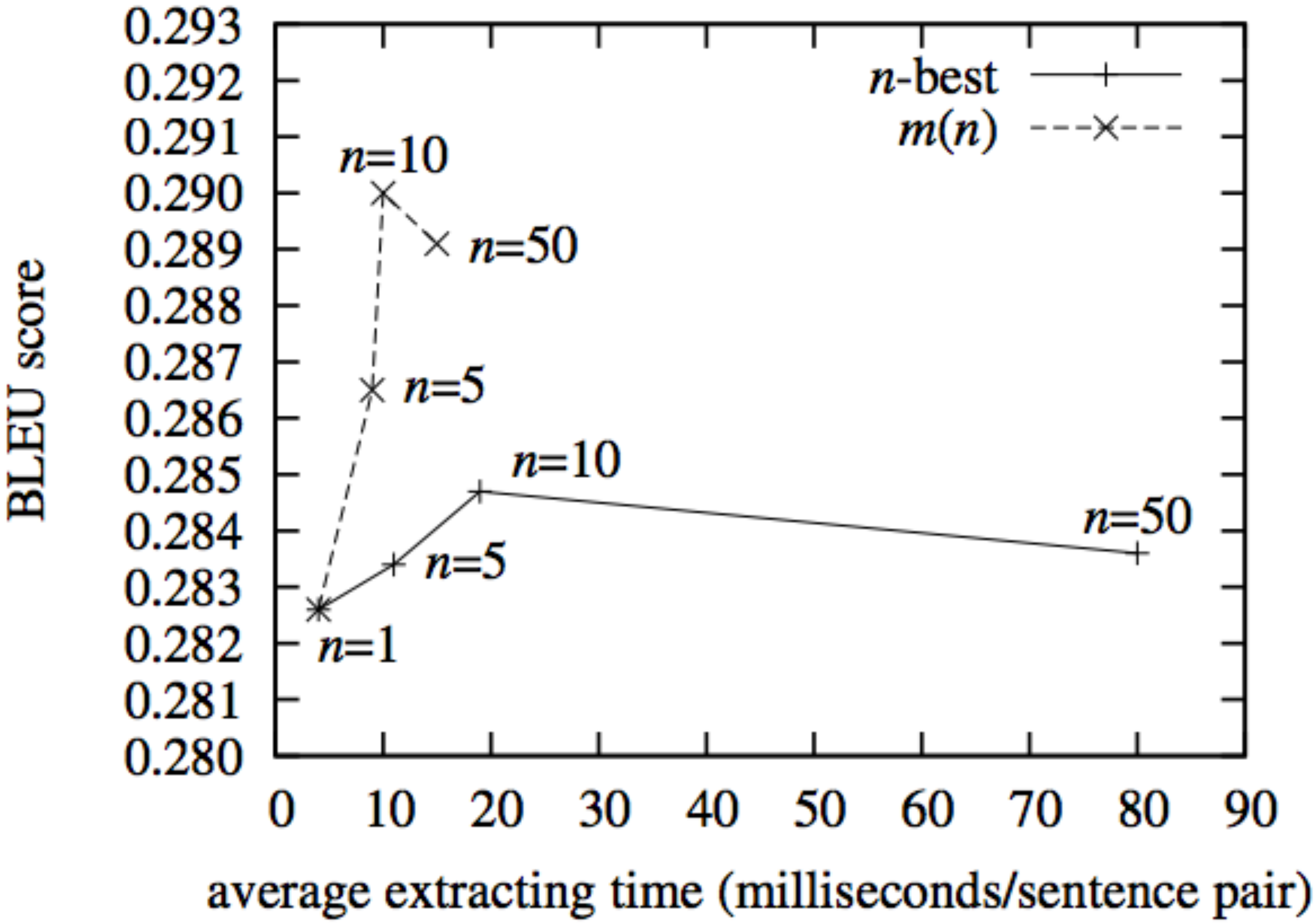
用表的技巧

无线、单线、双线的区别



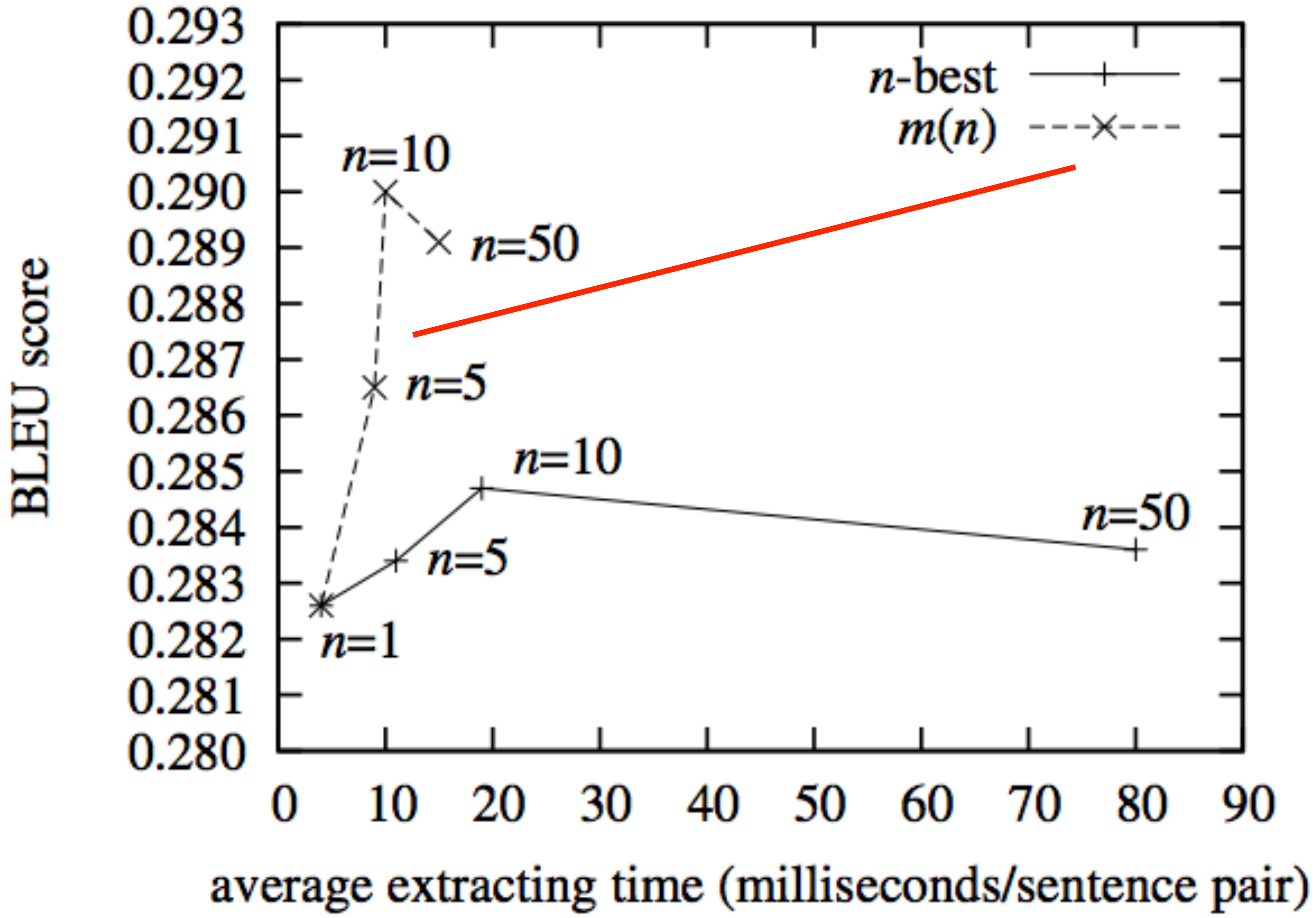
Method	Feature	MT02	MT03	MT04	MT05	MT06	MT08	All
RNNSEARCH	N/A	33.45	30.93	32.57	29.86	29.03	21.85	29.11
CPR	N/A	33.84	31.18	33.26	30.67	29.63	22.38	29.72
POSTREG	BD	34.65	31.53	33.82	30.66	29.81	22.55	29.97
	PT	34.56	31.32	33.89	30.70	29.84	22.62	29.99
	LR	34.39	31.41	34.19	30.80	29.82	22.85	30.14
	BD+PT	34.66	32.05	34.54	31.22	30.70	22.84	30.60
	BD+PT+LR	34.37	31.42	34.18	30.99	29.90	22.87	30.20
this work	BD	**36.61**	33.47	36.04	32.96	32.46	**24.78**	32.27
	PT	35.07	32.11	34.73	31.84	30.82	23.23	30.86
	CP	34.68	31.99	34.67	31.37	30.80	23.34	30.76
	LR	34.57	31.89	34.95	31.80	31.43	23.75	31.12
	BD+PT	36.30	**33.83**	36.02	32.98	32.53	24.54	32.29
	BD+PT+CP	36.11	33.64	36.36	**33.11**	32.53	24.57	32.39
	BD+PT+CP+LR	36.10	33.64	**36.48**	33.08	**32.90**	24.63	**32.51**

用图的误区

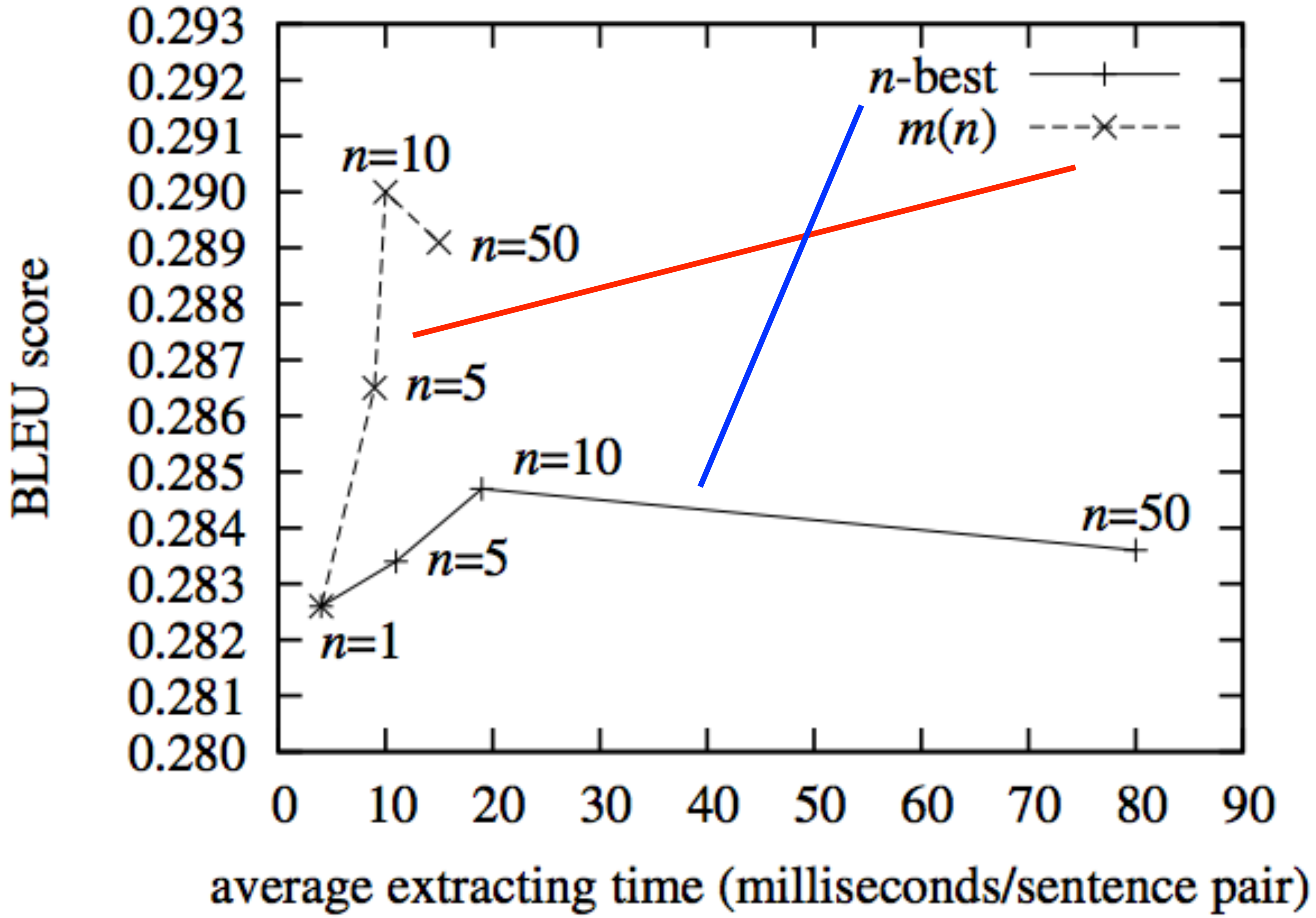


BLEU score
0.293
0.292
0.291
0.290
0.289
0.288
0.287
0.286
0.285
0.284
0.283
0.282
0.281
0.280
0
10
20
30
40
50
60
70
80
90
average extracting time (milliseconds/sentence pair)
n-best
m(n)
n=10
n=50
n=5
n=10
n=50
n=5
n=1

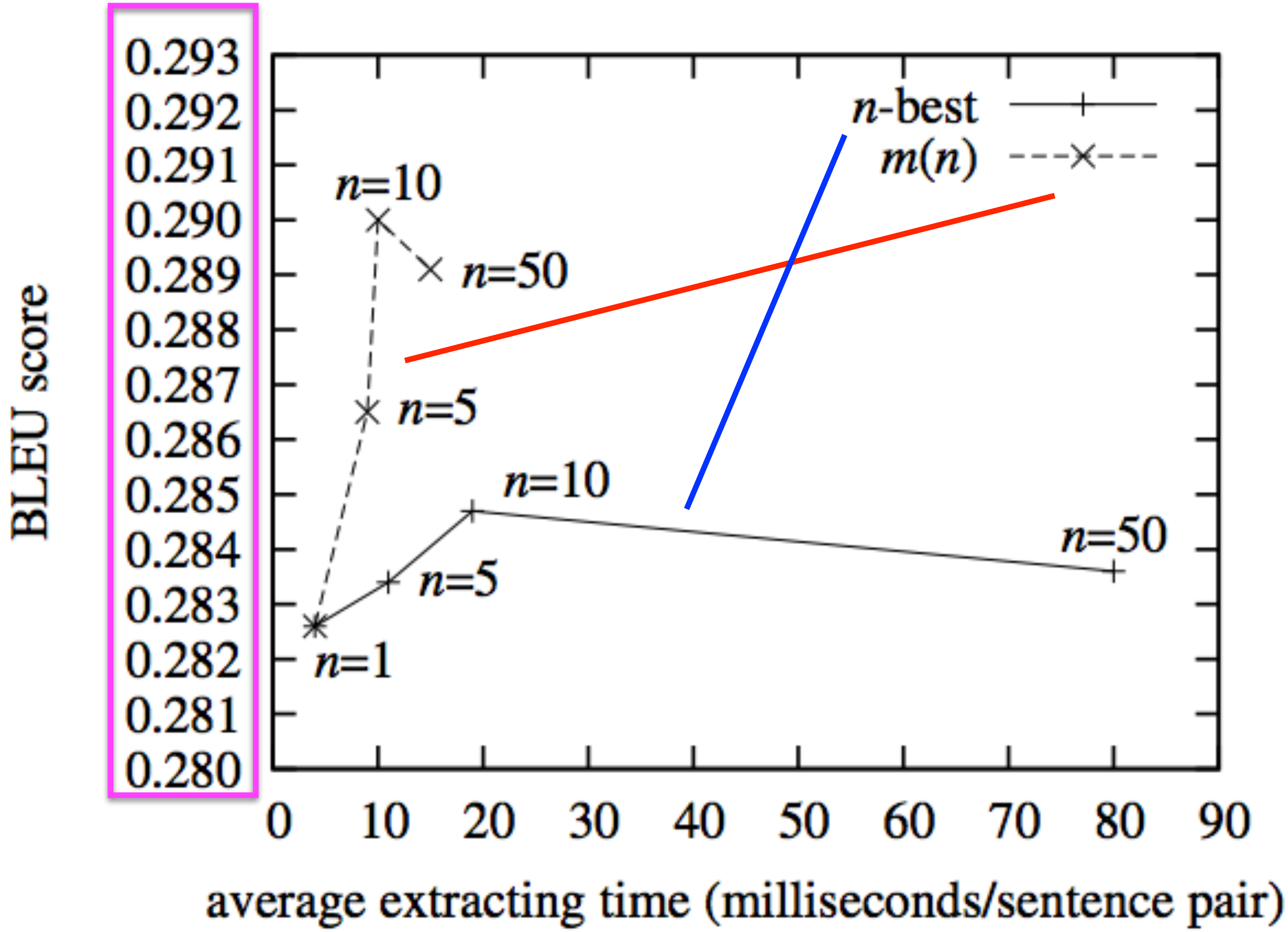
用图的误区


0.293
0.292
0.291
0.290
0.289
0.288
0.287
0.286
0.285
0.284
0.283
0.282
0.281
0.280
BLEU score
0
10
20
30
40
50
60
70
80
90
average extracting time (milliseconds/sentence pair)
n-best
m(n)
n=10
n=50
n=5
n=1
n=10
n=5
n=50

用图的误区


0.293
0.292
0.291
0.290
0.289
0.288
0.287
0.286
0.285
0.284
0.283
0.282
0.281
0.280
BLEU score
0
10
20
30
40
50
60
70
80
90
average extracting time (milliseconds/sentence pair)
n-best
m(n)
n=10
n=50
n=5
n=1
n=10
n=5
n=50

用图的误区


BLEU score
0.293
0.292
0.291
0.290
0.289
0.288
0.287
0.286
0.285
0.284
0.283
0.282
0.281
0.280
0
10
20
30
40
50
60
70
80
90
average extracting time (milliseconds/sentence pair)
n-best
m(n)
n=10
n=50
n=5
n=10
n=50
n=5
n=1

Caption包含充分的信息

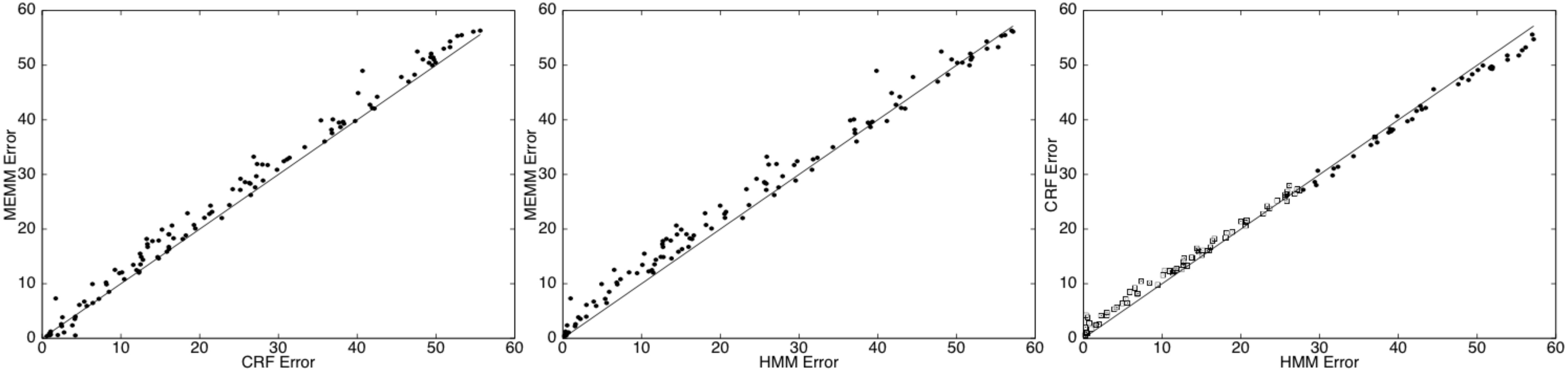


Figure 3. Plots of 2×2 error rates for HMMs, CRFs, and MEMMs on randomly generated synthetic data sets, as described in Section 5.2. As the data becomes "more second order," the error rates of the test models increase. As shown in the left plot, the CRF typically significantly outperforms the MEMM. The center plot shows that the HMM outperforms the MEMM. In the right plot, each open square represents a data set with $\alpha < \frac{1}{2}$, and a solid circle indicates a data set with $\alpha \geq \frac{1}{2}$. The plot shows that when the data is mostly second order ($\alpha \geq \frac{1}{2}$), the discriminatively trained CRF typically outperforms the HMM. These experiments are not designed to demonstrate the advantages of the additional representational power of CRFs and MEMMs relative to HMMs.

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

Caption包含充分的信息

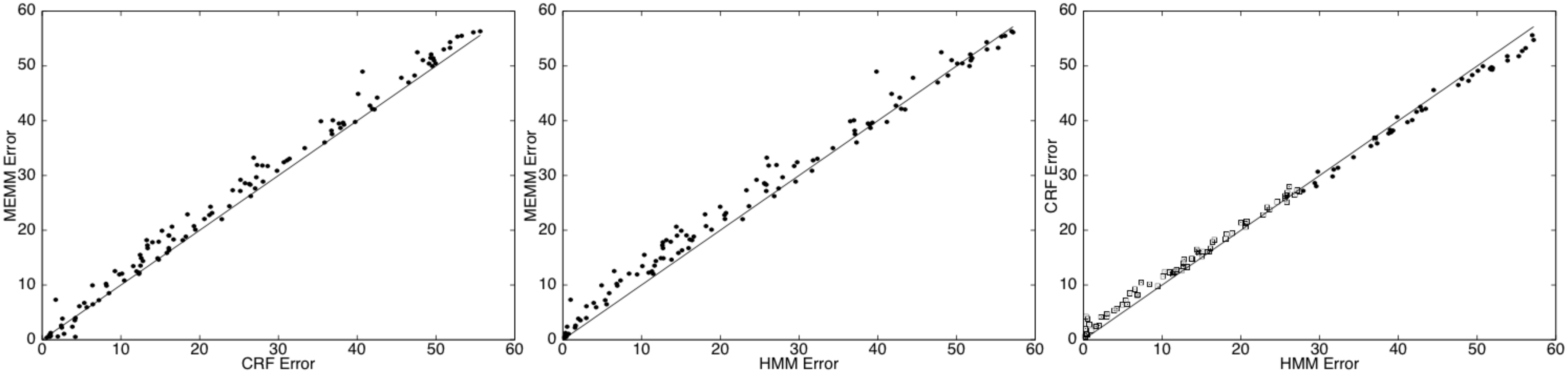


Figure 3. Plots of 2×2 error rates for HMMs, CRFs, and MEMMs on randomly generated synthetic data sets, as described in Section 5.2. As the data becomes "more second order," the error rates of the test models increase. As shown in the left plot, the CRF typically significantly outperforms the MEMM. The center plot shows that the HMM outperforms the MEMM. In the right plot, each open square represents a data set with $\alpha < \frac{1}{2}$, and a solid circle indicates a data set with $\alpha \geq \frac{1}{2}$. The plot shows that when the data is mostly second order ($\alpha \geq \frac{1}{2}$), the discriminatively trained CRF typically outperforms the HMM. These experiments are not designed to demonstrate the advantages of the additional representational power of CRFs and MEMMs relative to HMMs.

最好能直接看懂图，不用再去看正文

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

相关工作的写作技巧

如何写相关工作

错误

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

通过批评乃至攻击前人工作证明你的工作的创新性

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

通过批评乃至攻击前人工作证明你的工作的创新性

正确

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

通过批评乃至攻击前人工作证明你的工作的创新性

正确

向审稿人显示你对本领域具有全面深刻的把握

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

通过批评乃至攻击前人工作证明你的工作的创新性

正确

向审稿人显示你对本领域具有全面深刻的把握

通过与前人工作的对比凸显你的工作的创新性

如何写相关工作

错误

没有引用重要论文（可以直接作为rejection的理由）

简单的罗列和堆砌，缺乏深刻到位的评论

通过批评乃至攻击前人工作证明你的工作的创新性

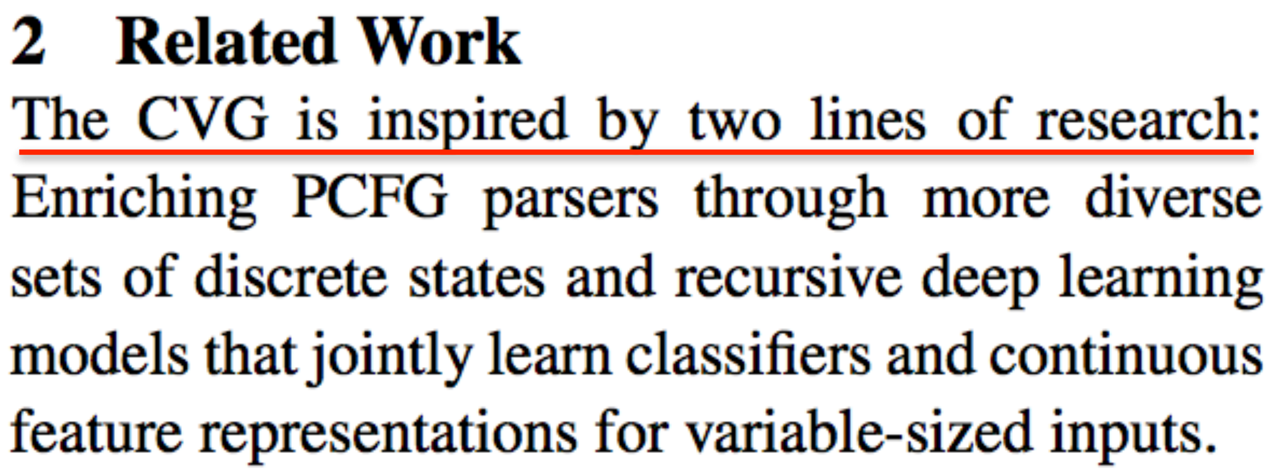
正确

向审稿人显示你对本领域具有全面深刻的把握

通过与前人工作的对比凸显你的工作的创新性

为读者梳理领域的发展脉络，获得全局的认识

例子

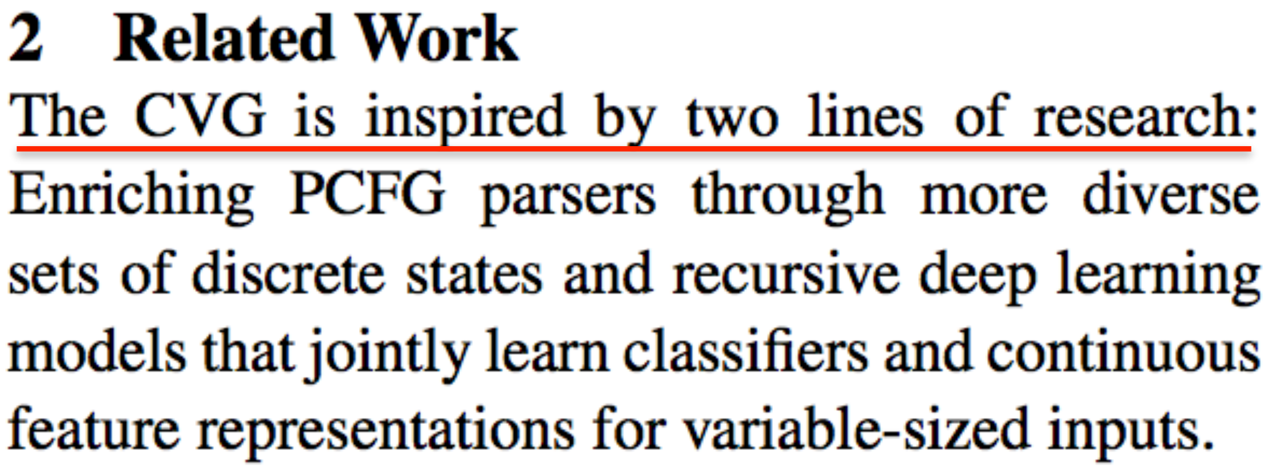


2 Related Work

The CVG is inspired by two lines of research: Enriching PCFG parsers through more diverse sets of discrete states and recursive deep learning models that jointly learn classifiers and continuous feature representations for variable-sized inputs.

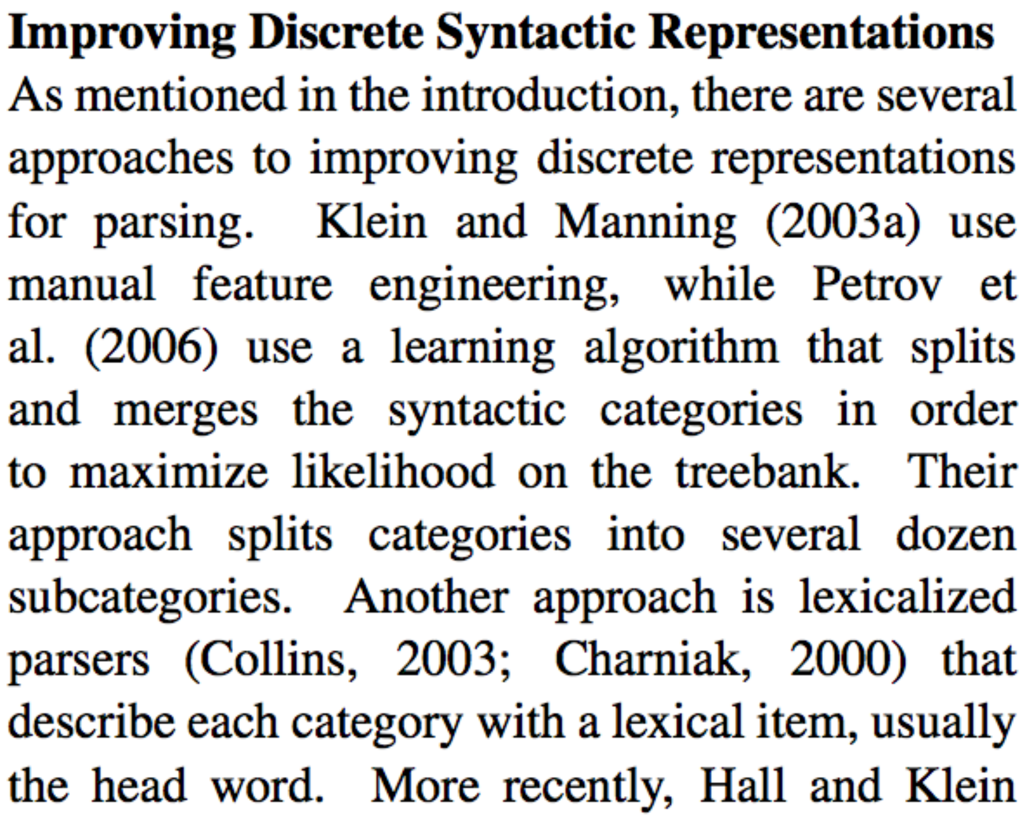
Richard Socher, John Bauer, Christopher Manning, and Andrew Ng. **Parsing with Compositional Vector Grammars**. In *ACL 2013*.

例子



2 Related Work

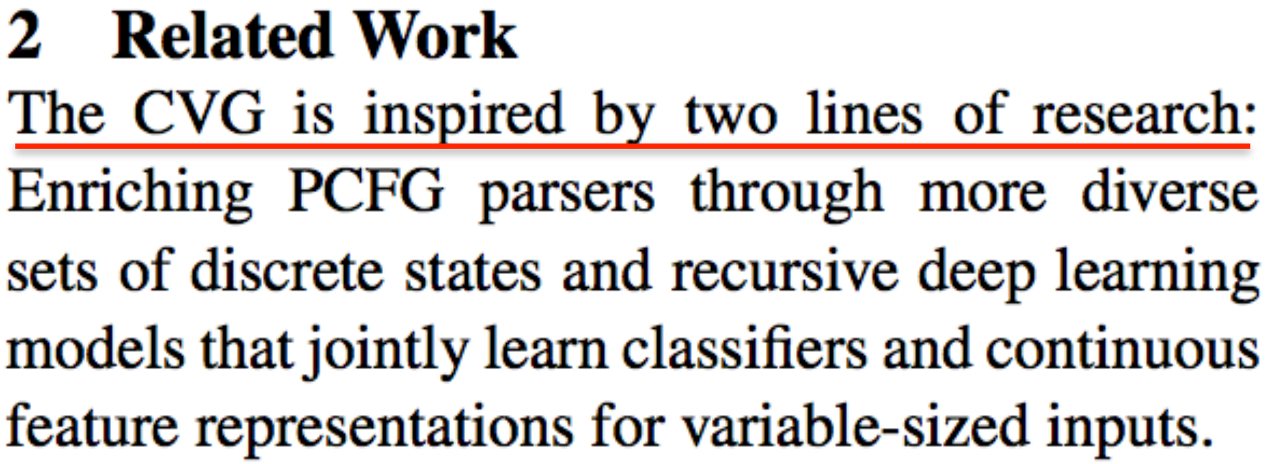
The CVG is inspired by two lines of research: Enriching PCFG parsers through more diverse sets of discrete states and recursive deep learning models that jointly learn classifiers and continuous feature representations for variable-sized inputs.



Improving Discrete Syntactic Representations As mentioned in the introduction, there are several approaches to improving discrete representations for parsing. Klein and Manning (2003a) use manual feature engineering, while Petrov et al. (2006) use a learning algorithm that splits and merges the syntactic categories in order to maximize likelihood on the treebank. Their approach splits categories into several dozen subcategories. Another approach is lexicalized parsers (Collins, 2003; Charniak, 2000) that describe each category with a lexical item, usually the head word. More recently, Hall and Klein

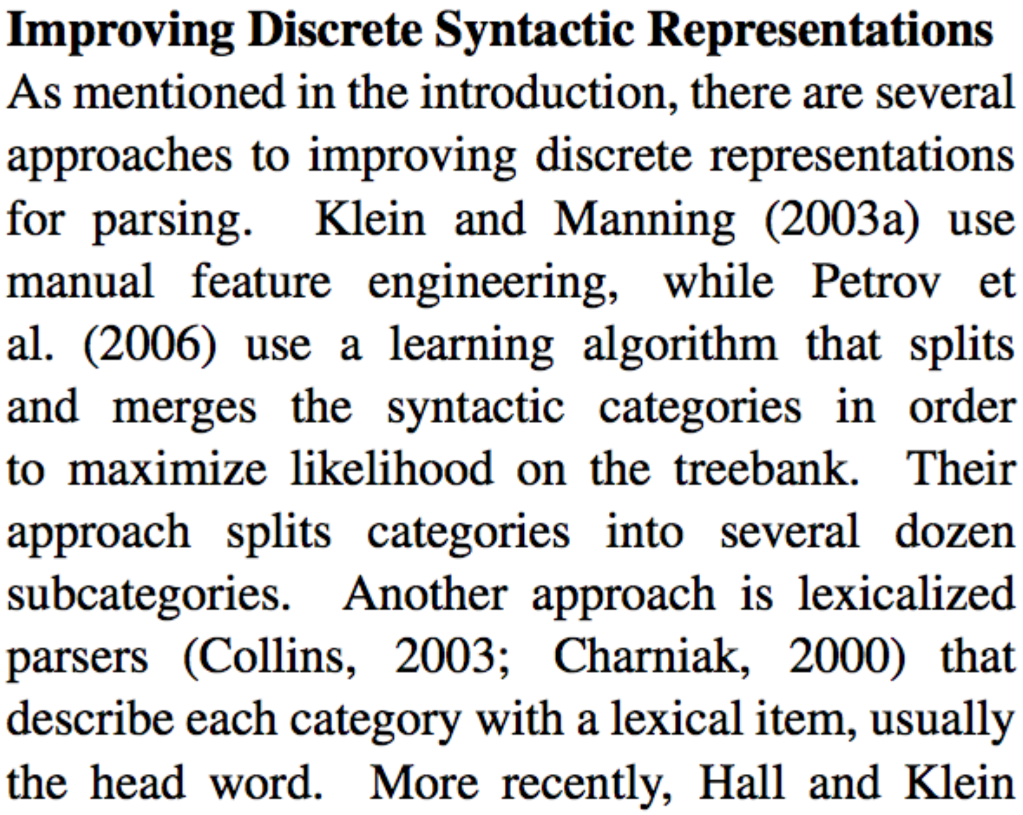
Richard Socher, John Bauer, Christopher Manning, and Andrew Ng. **Parsing with Compositional Vector Grammars**. In *ACL 2013*.

例子

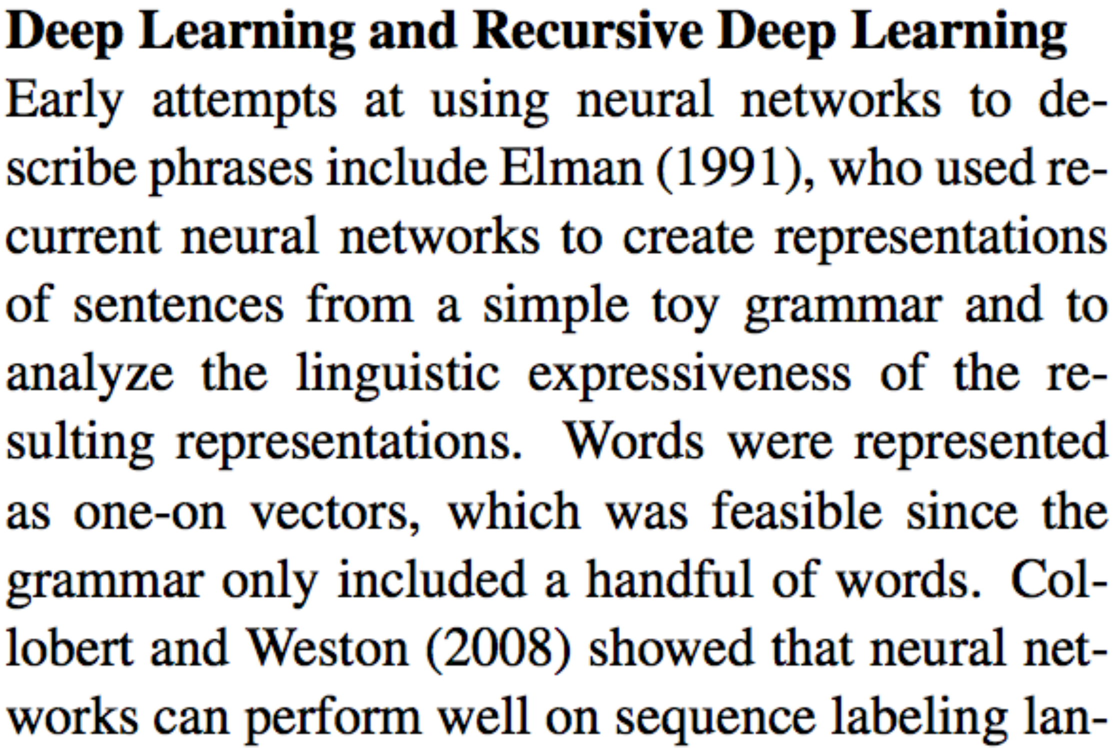


2 Related Work

The CVG is inspired by two lines of research: Enriching PCFG parsers through more diverse sets of discrete states and recursive deep learning models that jointly learn classifiers and continuous feature representations for variable-sized inputs.



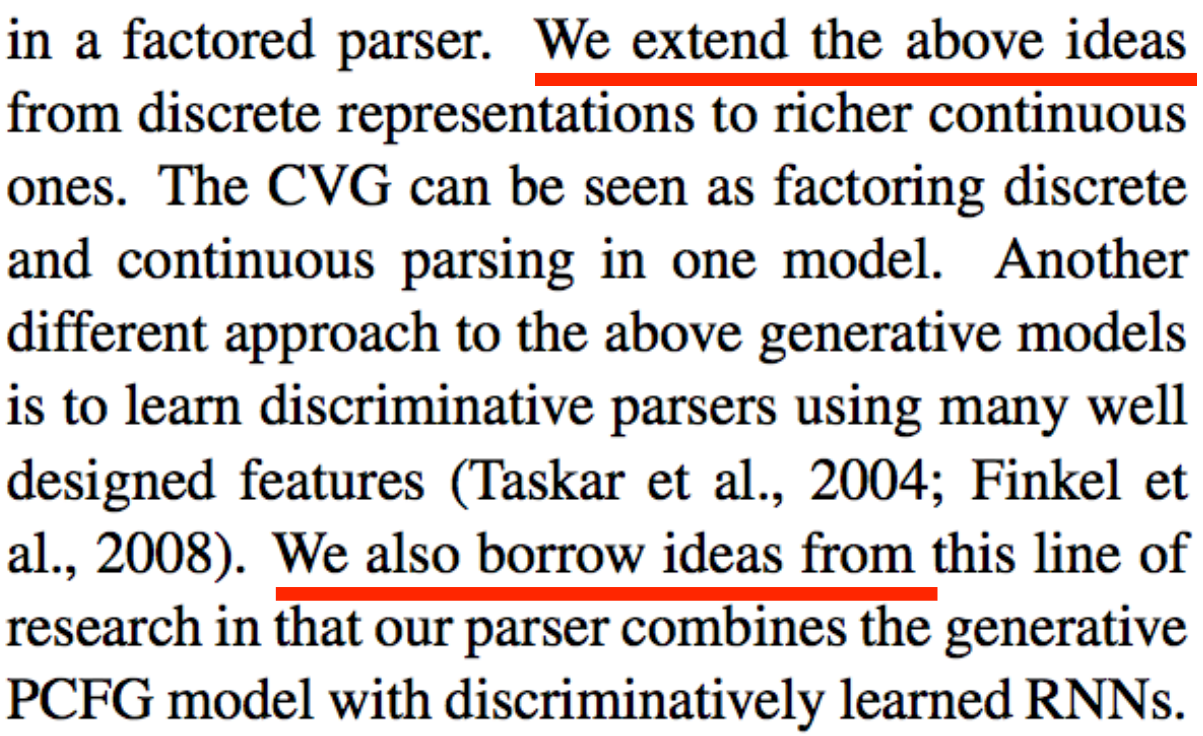
Improving Discrete Syntactic Representations
As mentioned in the introduction, there are several approaches to improving discrete representations for parsing. Klein and Manning (2003a) use manual feature engineering, while Petrov et al. (2006) use a learning algorithm that splits and merges the syntactic categories in order to maximize likelihood on the treebank. Their approach splits categories into several dozen subcategories. Another approach is lexicalized parsers (Collins, 2003; Charniak, 2000) that describe each category with a lexical item, usually the head word. More recently, Hall and Klein



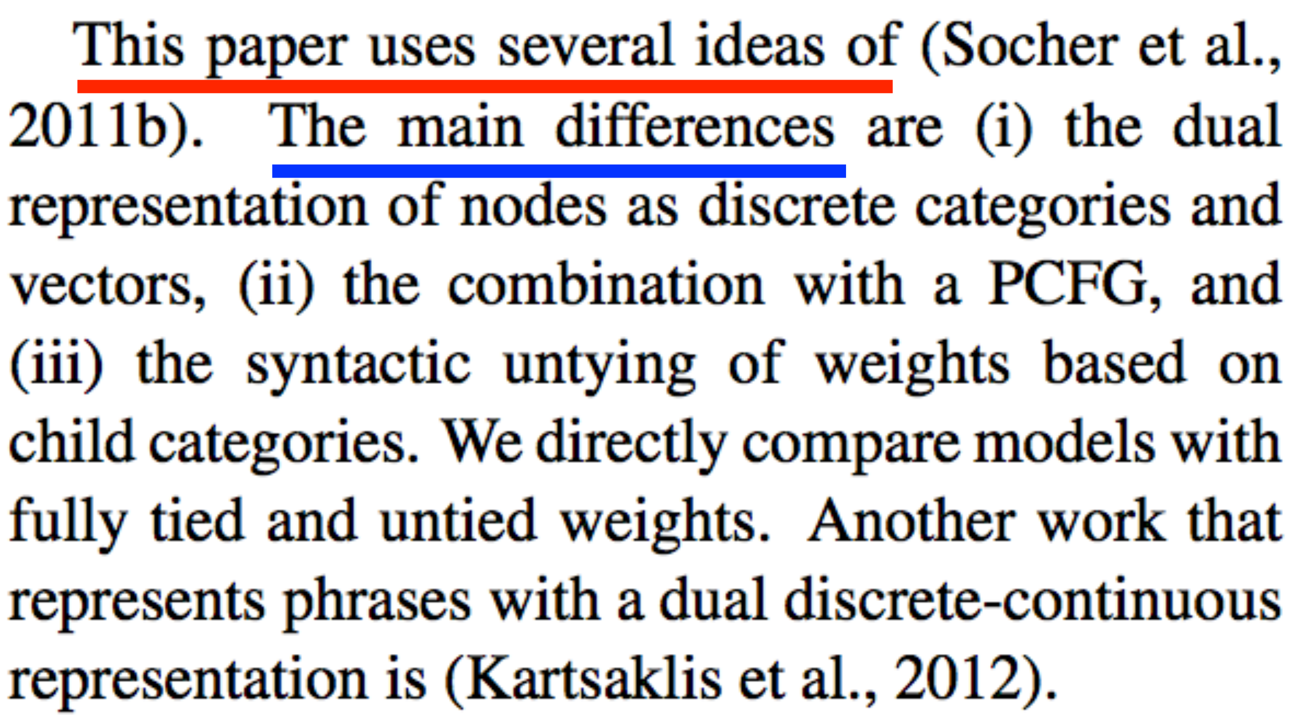
Deep Learning and Recursive Deep Learning
Early attempts at using neural networks to describe phrases include Elman (1991), who used recurrent neural networks to create representations of sentences from a simple toy grammar and to analyze the linguistic expressiveness of the resulting representations. Words were represented as one-on vectors, which was feasible since the grammar only included a handful of words. Collobert and Weston (2008) showed that neural networks can perform well on sequence labeling lan-

Richard Socher, John Bauer, Christopher Manning, and Andrew Ng. **Parsing with Compositional Vector Grammars**. In *ACL 2013*.

传承与创新



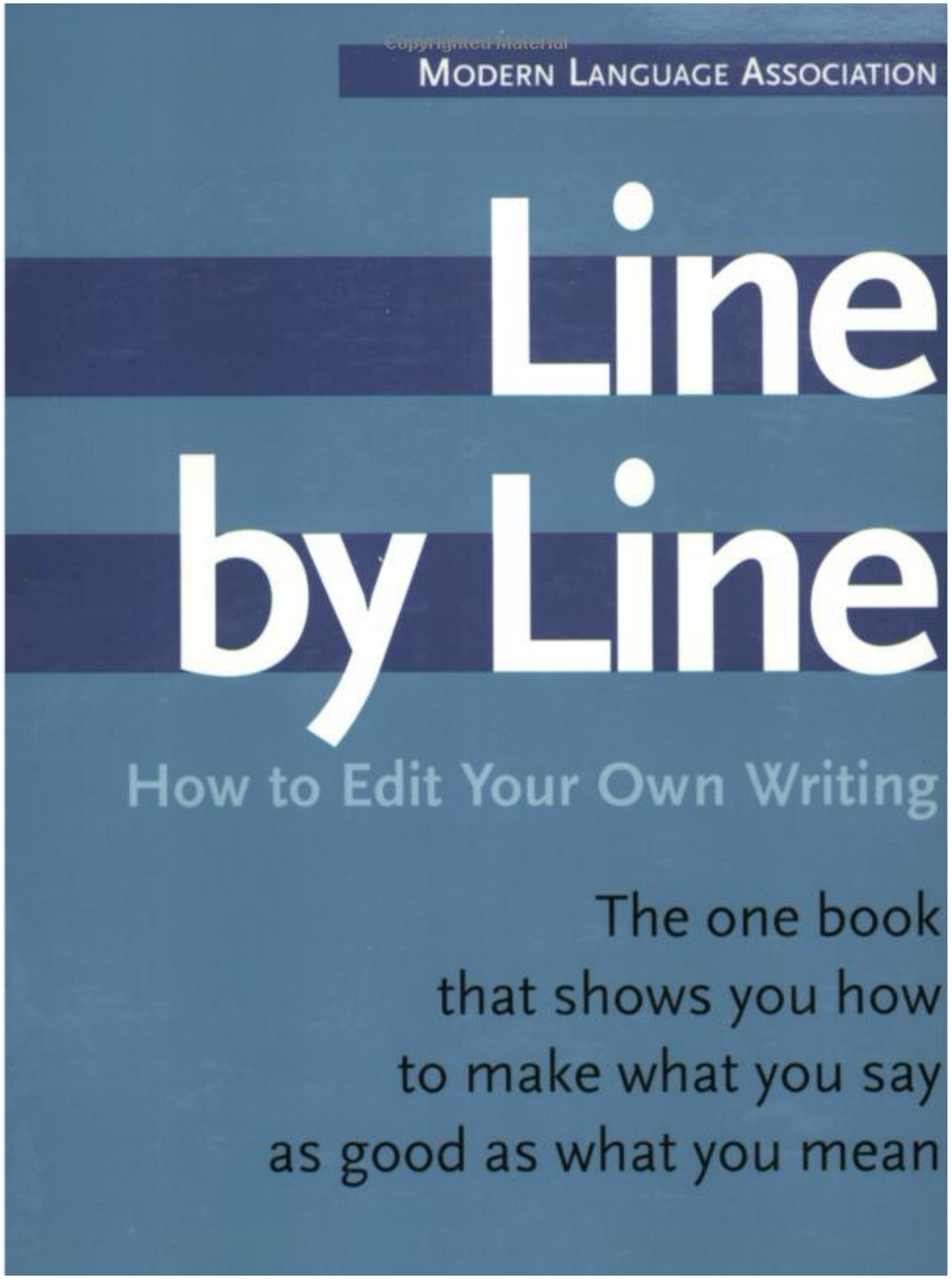
in a factored parser. We extend the above ideas from discrete representations to richer continuous ones. The CVG can be seen as factoring discrete and continuous parsing in one model. Another different approach to the above generative models is to learn discriminative parsers using many well designed features (Taskar et al., 2004; Finkel et al., 2008). We also borrow ideas from this line of research in that our parser combines the generative PCFG model with discriminatively learned RNNs.



This paper uses several ideas of (Socher et al., 2011b). The main differences are (i) the dual representation of nodes as discrete categories and vectors, (ii) the combination with a PCFG, and (iii) the syntactic untying of weights based on child categories. We directly compare models with fully tied and untied weights. Another work that represents phrases with a dual discrete-continuous representation is (Kartsaklis et al., 2012).

Richard Socher, John Bauer, Christopher Manning, and Andrew Ng. **Parsing with Compositional Vector Grammars**. In *ACL 2013*.

英文写作进阶



MODERN LANGUAGE ASSOCIATION

Line by Line

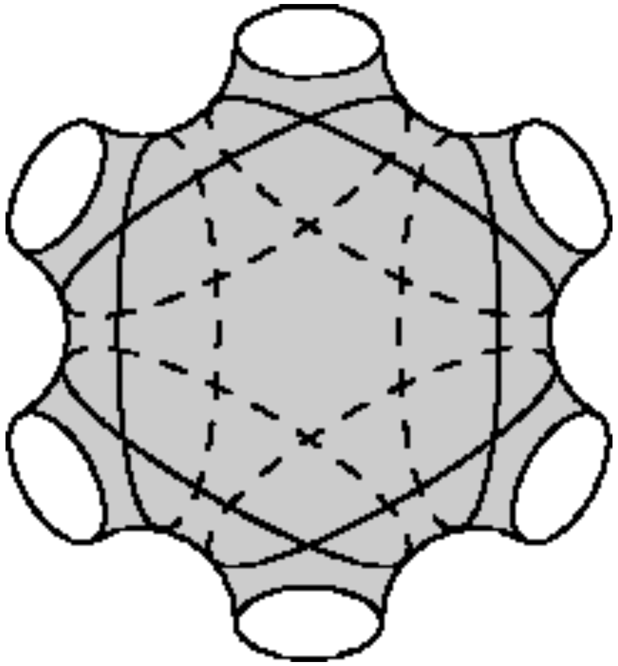
How to Edit Your Own Writing

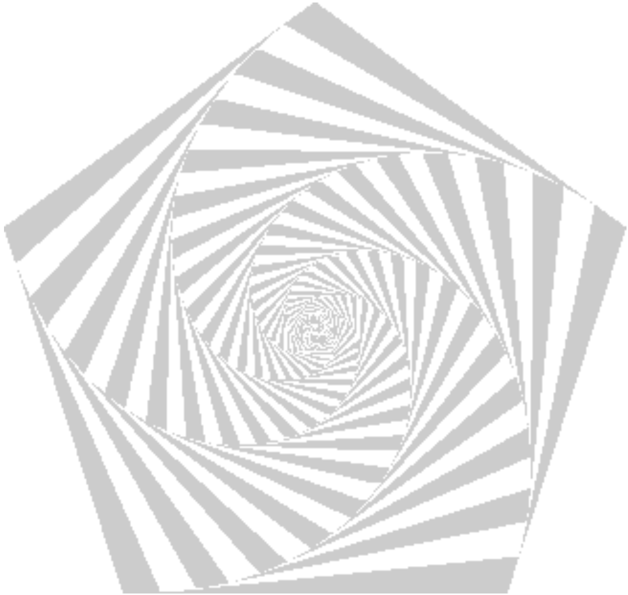
The one book
that shows you how
to make what you say
as good as what you mean

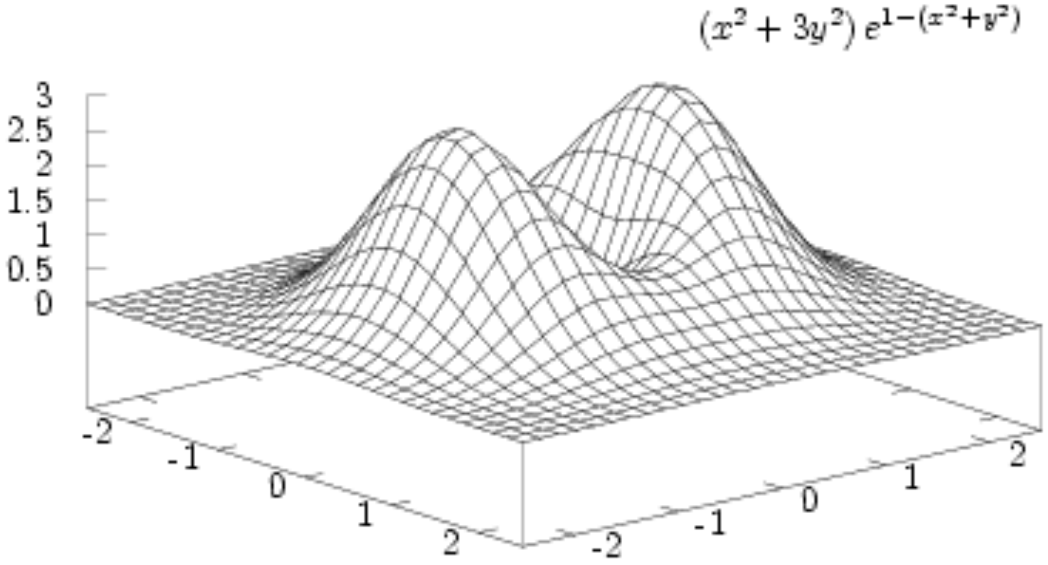
必须掌握的工具

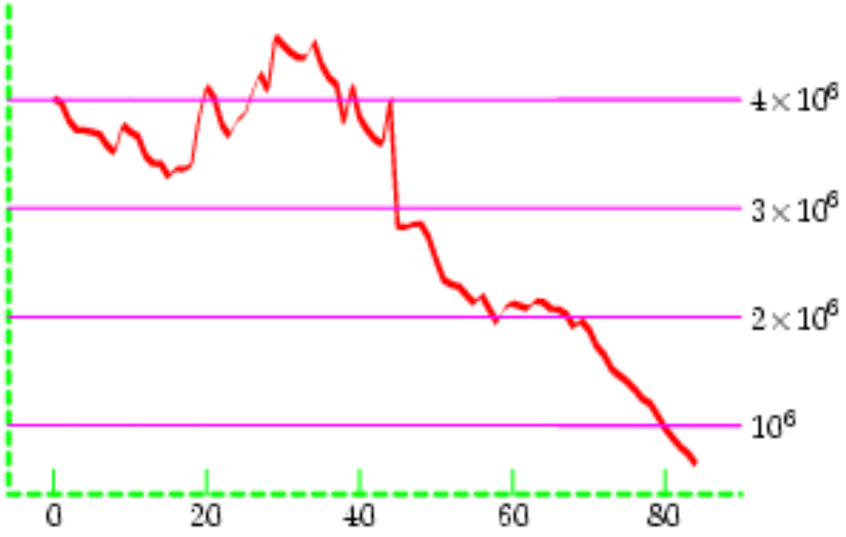
- LaTex
    - 强烈建议用LaTex代替Word
    - http://www.ctex.org/HomePage
- Bibtex
    - 自动生成参考文献列表
- MetaPost
    - 编程画矢量图

MetaPost






$(x^2 + 3y^2)\, e^{1-(x^2+y^2)}$
3
2.5
2
1.5
1
0.5
0
-2
-1
0
1
2
-2
-1
0
1
2


4×10^6
3×10^6
2×10^6
10^6
0
20
40
60
80

时间管理和获得反馈

- coarse-to-fine
  - 截稿前一个月开始写
  - 每隔两天改一次
- 听取不同背景读者的反馈意见
  - 专家：专业意见
  - 非专家：发现信息壁垒
- 写到极致，完成完美精致的艺术品

平时如何学习写论文

- 研读和剖析公认的经典范文，学习写作技巧，“模拟写作”
- 平时常做研究笔记，多动笔头
- 认真做好组会报告，练习和提高表达能力
- 在投稿过程中多听取导师、同学和审稿人的意见

论文写作的境界

- **初学乍练**：按照自己的想法写，不学习范文
- **初窥门径**：掌握了范文技巧，形式上较规范
- **融会贯通**：抛开范文，按照自己的想法写
- **炉火纯青**：开拓创新，引领写作技巧新潮流
- **登峰造极**：注重深层思想，而非表层形式

论文不仅仅是写作

- **视野**（Vision）：把握脉络，捕捉战机
- **品味**（Taste）：选择的智慧
- **态度**（Attitude）：治学严谨
- **技能**（Skills）：严格训练，熟能生巧

总结

- 写论文本质是分享思想，呈现信息
- 信息的呈现符合读者的认知惯性
- 全心全意为读者服务，降低阅读难度，提高愉悦感
- 细节决定成败
- 不要本末倒置：创新至上，技法为辅。

谢谢

审稿

审稿

你以为审稿人应该是这样审稿的：

审稿

你以为审稿人应该是这样审稿的：

审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个border line或以上的分数。

审稿

你以为审稿人应该是这样审稿的：

审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个border line或以上的分数。

审稿人实际上往往是这样审稿的：

审稿

你以为审稿人应该是这样审稿的：

审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个border line或以上的分数。

审稿人实际上往往是这样审稿的：

他不一定是专家，一直忙于其他事，在deadline到来之前一天要完成n篇。审稿时他往往先看题目、摘要，扫一下introduction（知道你做什么），然后直接翻到最后找核心实验结果（做得好不好），然后基本确定录还是不录（也许只用5分钟！）。如果决定录，剩下就是写些赞美的话，指出些次要的小毛病。如果决定拒，下面的过程就是细看中间部分找理由拒了。

审稿

你以为审稿人应该是这样审稿的：

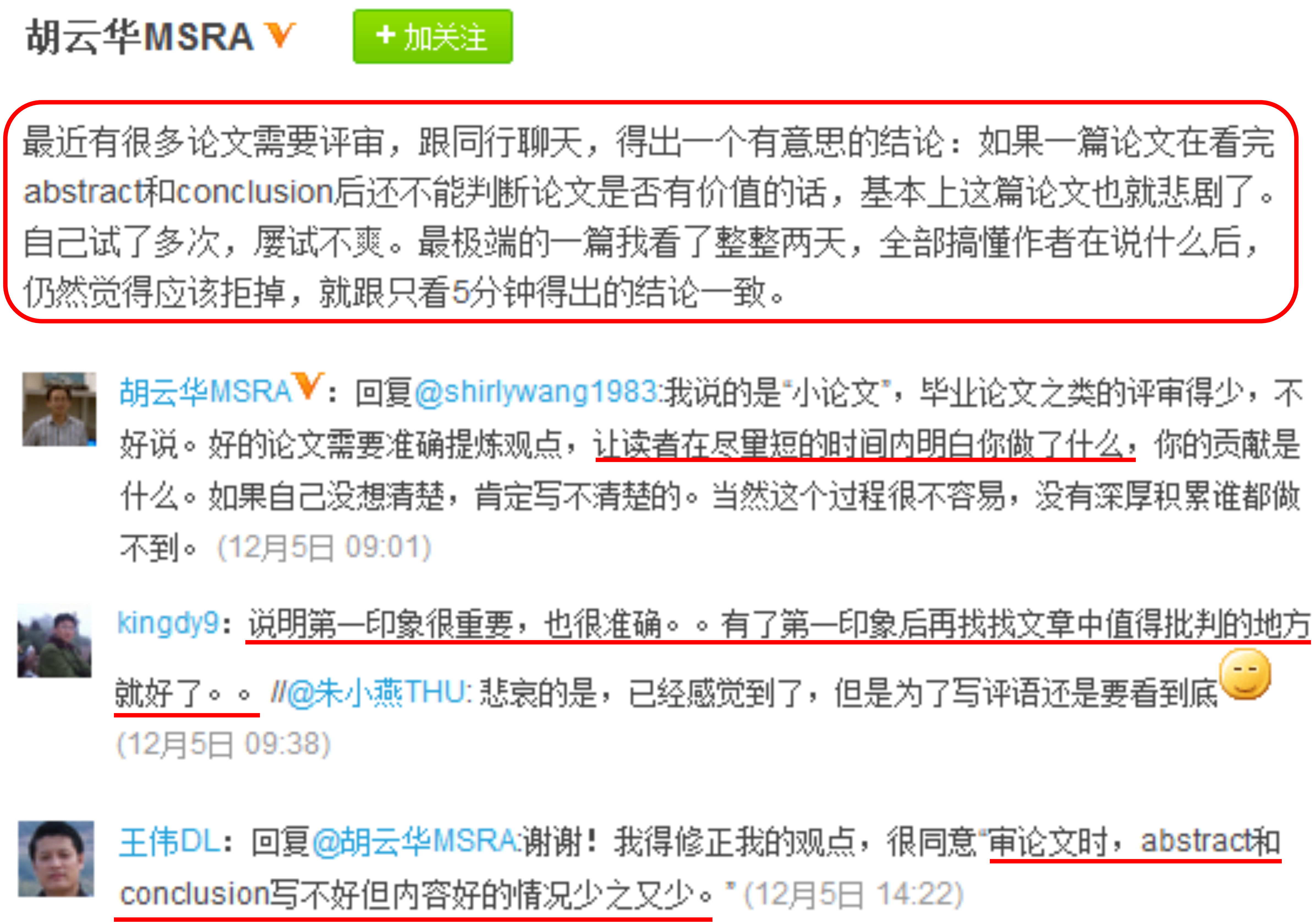
审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个border line或以上的分数。

审稿人实际上往往是这样审稿的：

他不一定是专家，一直忙于其他事，在deadline到来之前一天要完成n篇。审稿时他往往先看题目、摘要，扫一下introduction（知道你做什么），然后直接翻到最后找核心实验结果（做得好不好），然后基本确定录还是不录（也许只用5分钟！）。如果决定录，剩下就是写些赞美的话，指出些次要的小毛病。如果决定拒，下面的过程就是细看中间部分找理由拒了。

第一印象定录拒，5分钟内打动审稿人

微博上的佐证



胡云华MSRA V

最近有很多论文需要评审，跟同行聊天，得出一个有意思的结论：如果一篇论文在看完abstract和conclusion后还不能判断论文是否有价值的话，基本上这篇论文也就悲剧了。自己试了多次，屡试不爽。最极端的一篇我看了整整两天，全部搞懂作者在说什么后，仍然觉得应该拒掉，就跟只看5分钟得出的结论一致。

胡云华MSRA V：回复@shirlywang1983:我说的是“小论文”，毕业论文之类的评审得少，不好说。好的论文需要准确提炼观点，让读者在尽量短的时间内明白你做了什么，你的贡献是什么。如果自己没想清楚，肯定写不清楚的。当然这个过程很不容易，没有深厚积累谁都做不到。(12月5日 09:01)

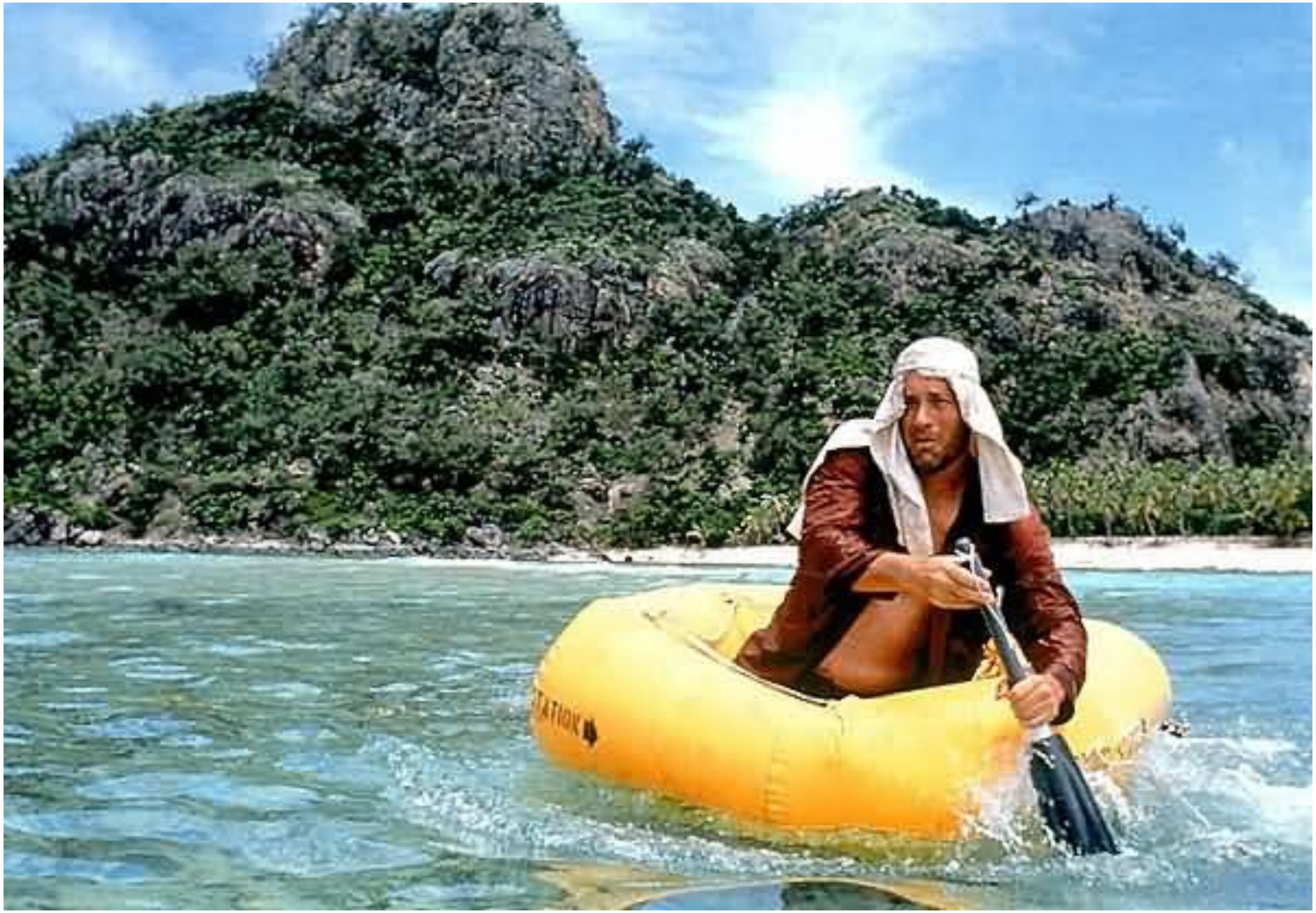
kingdy9：说明第一印象很重要，也很准确。。有了第一印象后再找找文章中值得批判的地方就好了。。 //@朱小燕THU:悲哀的是，已经感觉到了，但是为了写评语还是要看到底 (12月5日 09:38)

王伟DL：回复@胡云华MSRA:谢谢！我得修正我的观点，很同意“审论文时，abstract和conclusion写不好但内容好的情况少之又少。”(12月5日 14:22)

如何选择方向?

选择热门的方向

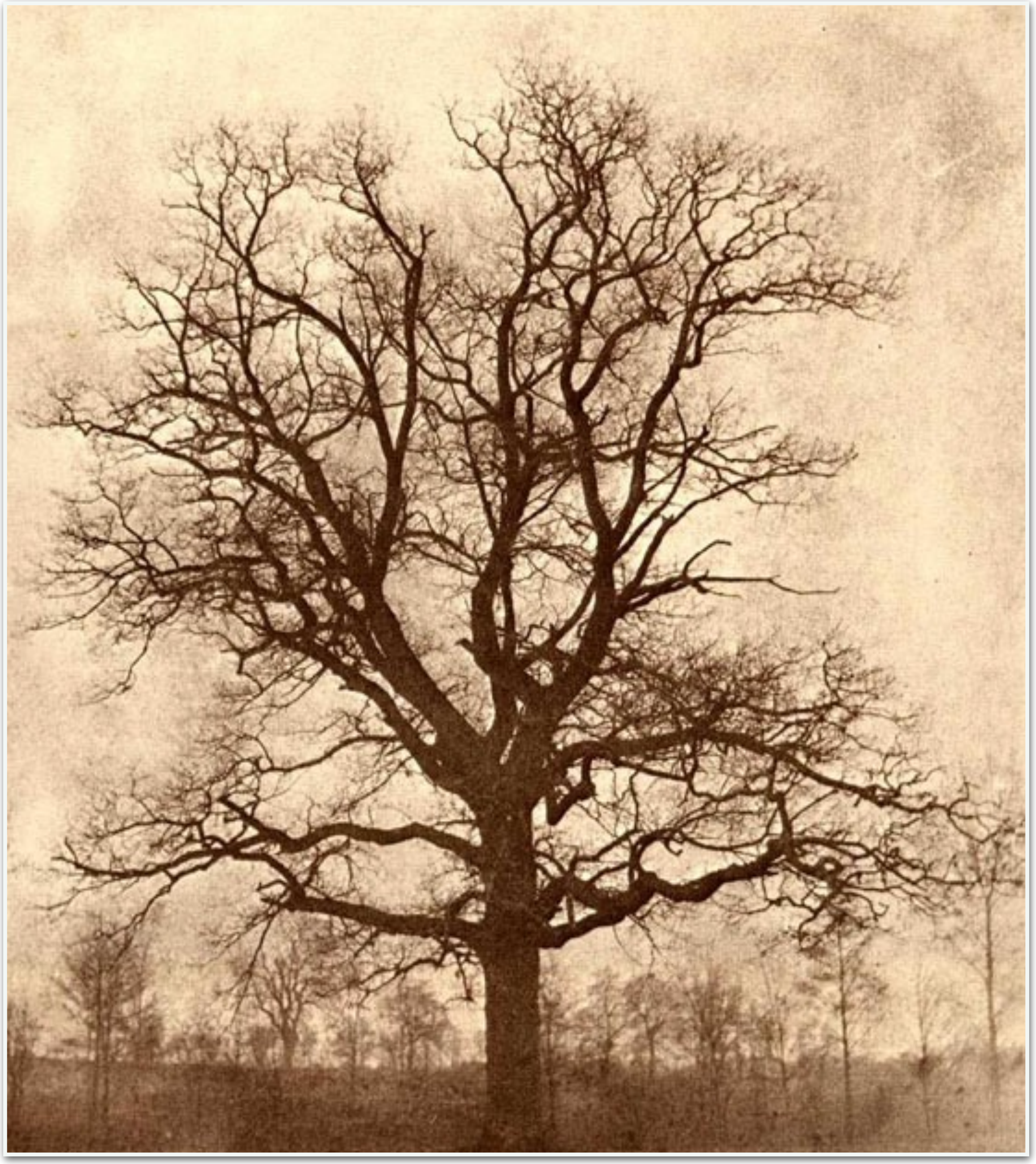
选择冷门的方向



“It is not worth an intelligent man's time to be in the majority. ***By definition****, there are already enough people to do that.”*

--- G. H. Hardy (1877-1947)

选择的智慧



选择的智慧

重要问题、重大挑战

选择的智慧

重要问题、重大挑战

自己感兴趣

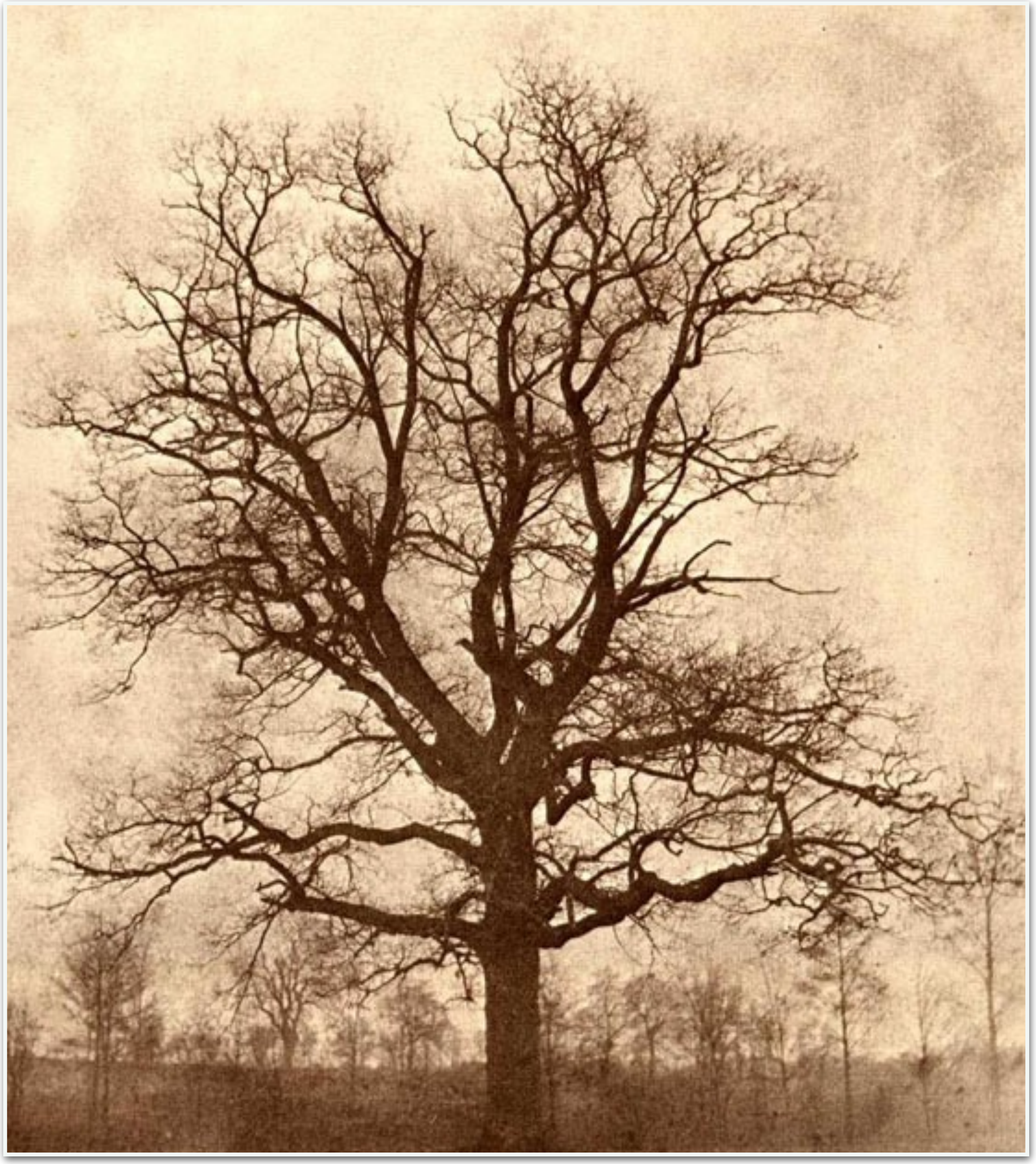
选择的智慧

重要问题、重大挑战

自己感兴趣

即将成为热门

选择的智慧



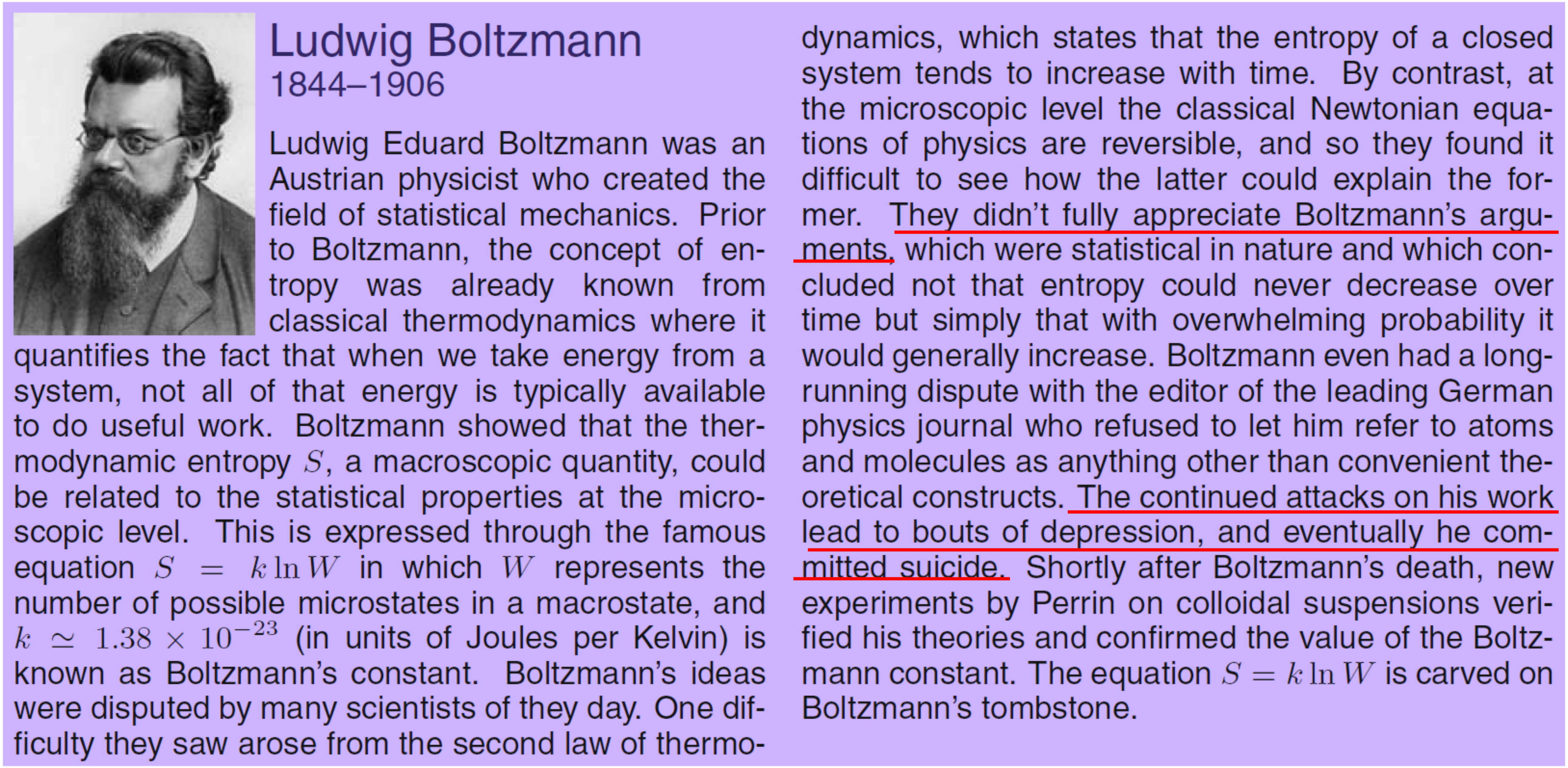
重要问题、重大挑战

自己感兴趣

即将成为热门

高风险性

做好不被承认的准备

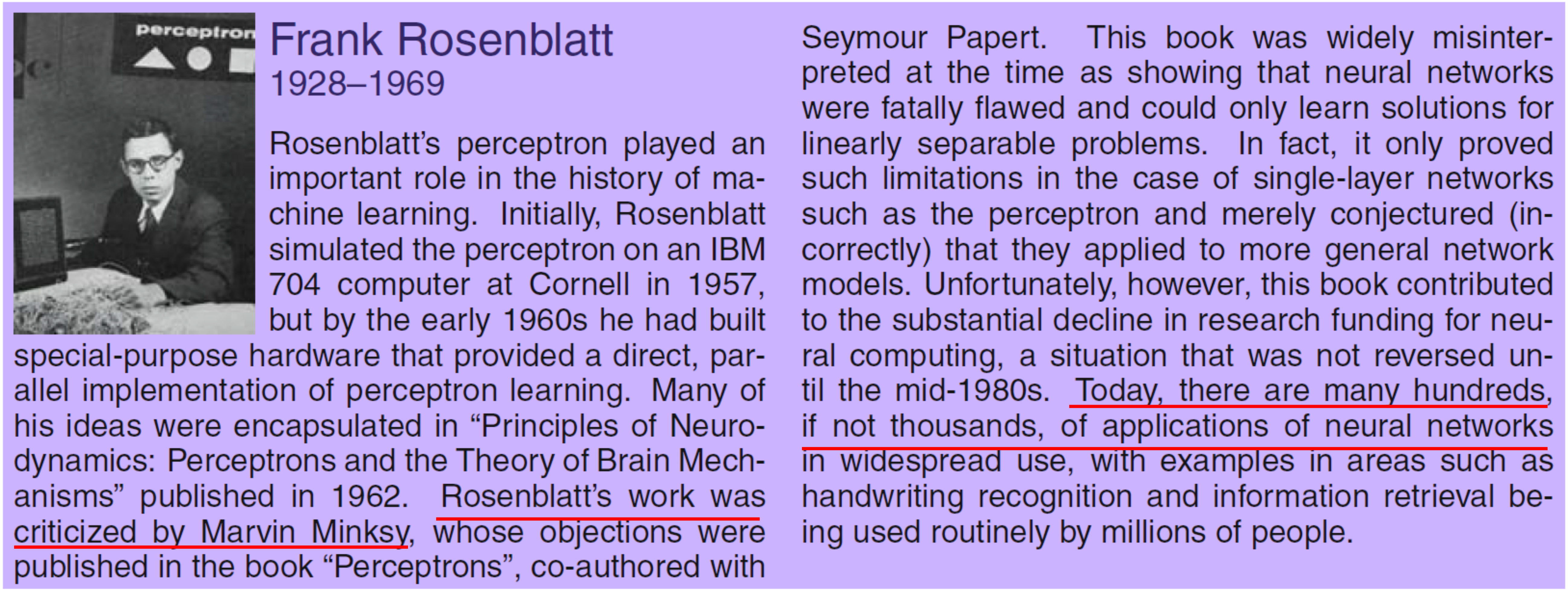


Ludwig Boltzmann
1844–1906

Ludwig Eduard Boltzmann was an Austrian physicist who created the field of statistical mechanics. Prior to Boltzmann, the concept of entropy was already known from classical thermodynamics where it quantifies the fact that when we take energy from a system, not all of that energy is typically available to do useful work. Boltzmann showed that the thermodynamic entropy S, a macroscopic quantity, could be related to the statistical properties at the microscopic level. This is expressed through the famous equation $S = k \ln W$ in which W represents the number of possible microstates in a macrostate, and $k \simeq 1.38 \times 10^{-23}$ (in units of Joules per Kelvin) is known as Boltzmann's constant. Boltzmann's ideas were disputed by many scientists of they day. One difficulty they saw arose from the second law of thermodynamics, which states that the entropy of a closed system tends to increase with time. By contrast, at the microscopic level the classical Newtonian equations of physics are reversible, and so they found it difficult to see how the latter could explain the former. They didn't fully appreciate Boltzmann's arguments, which were statistical in nature and which concluded not that entropy could never decrease over time but simply that with overwhelming probability it would generally increase. Boltzmann even had a long-running dispute with the editor of the leading German physics journal who refused to let him refer to atoms and molecules as anything other than convenient theoretical constructs. The continued attacks on his work lead to bouts of depression, and eventually he committed suicide. Shortly after Boltzmann's death, new experiments by Perrin on colloidal suspensions verified his theories and confirmed the value of the Boltzmann constant. The equation $S = k \ln W$ is carved on Boltzmann's tombstone.

Pattern Recognition and Machine Learning, C. Bishop

做好不被承认的准备



Frank Rosenblatt
1928–1969

Rosenblatt's perceptron played an important role in the history of machine learning. Initially, Rosenblatt simulated the perceptron on an IBM 704 computer at Cornell in 1957, but by the early 1960s he had built special-purpose hardware that provided a direct, parallel implementation of perceptron learning. Many of his ideas were encapsulated in "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms" published in 1962. Rosenblatt's work was criticized by Marvin Minksy, whose objections were published in the book "Perceptrons", co-authored with Seymour Papert. This book was widely misinterpreted at the time as showing that neural networks were fatally flawed and could only learn solutions for linearly separable problems. In fact, it only proved such limitations in the case of single-layer networks such as the perceptron and merely conjectured (incorrectly) that they applied to more general network models. Unfortunately, however, this book contributed to the substantial decline in research funding for neural computing, a situation that was not reversed until the mid-1980s. Today, there are many hundreds, if not thousands, of applications of neural networks in widespread use, with examples in areas such as handwriting recognition and information retrieval being used routinely by millions of people.

Pattern Recognition and Machine Learning, C. Bishop

像外行一样思考，像内行一样实践

Broadview
www.broadview.com.cn

像外行一样思考，像专家一样实践
——科研成功之道

素人のように
考え、
玄人として
実行する

向创新发出邀请！

【[日]金出武雄 著 马金成 王国强 译】

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

金出武雄

外行与内行

思考	实践	境界
外行	专家	独树一帜、炉火纯青
专家	专家	经验丰富、难脱窠臼
外行	外行	天马行空、眼高手低
专家	外行	思维僵化、束手无策

解决问题

解决问题

思维独立性

解决问题

思维独立性

先思考，再去查文献相互印证

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

符合读者的直觉：出乎意料，情理之中

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

符合读者的直觉：出乎意料，情理之中

数学意义

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

符合读者的直觉：出乎意料，情理之中

数学意义

使用数学工具做形式化，不臆造数学公式

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

符合读者的直觉：出乎意料，情理之中

数学意义

使用数学工具做形式化，不臆造数学公式

简洁优美

解决问题

思维独立性

先思考，再去查文献相互印证

符合直觉

符合读者的直觉：出乎意料，情理之中

数学意义

使用数学工具做形式化，不臆造数学公式

简洁优美

简单、干净、优美

其它

- 论文中每个数学符号都应当找得到定义，除非众所周知。永远不要不加说明就使用数学符号。
- 要避免数学符号冲突，使用符号列表
- 不要生造术语，尤其是中式译法，尽量符合惯例
- 集成所有信息元素，排版美观和专业

论文组织技巧

降低信息理解难度是关键

降低信息理解难度是关键

1 介绍
2 相关工作
3 方法
4 实验
5 结论

降低信息理解难度是关键

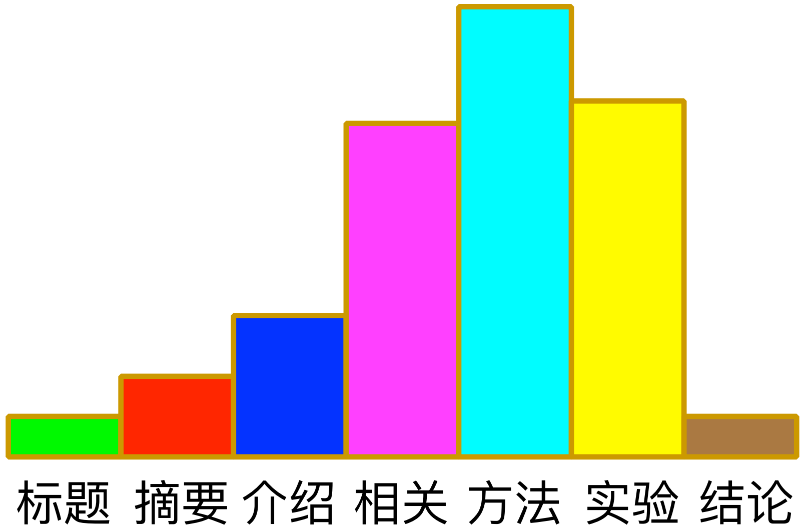
1 介绍
2 相关工作
3 方法
4 实验
5 结论

1 介绍
2 背景
3 方法
4 实验
5 相关工作
6 结论

降低信息理解难度是关键

1 介绍
2 相关工作
3 方法
4 实验
5 结论

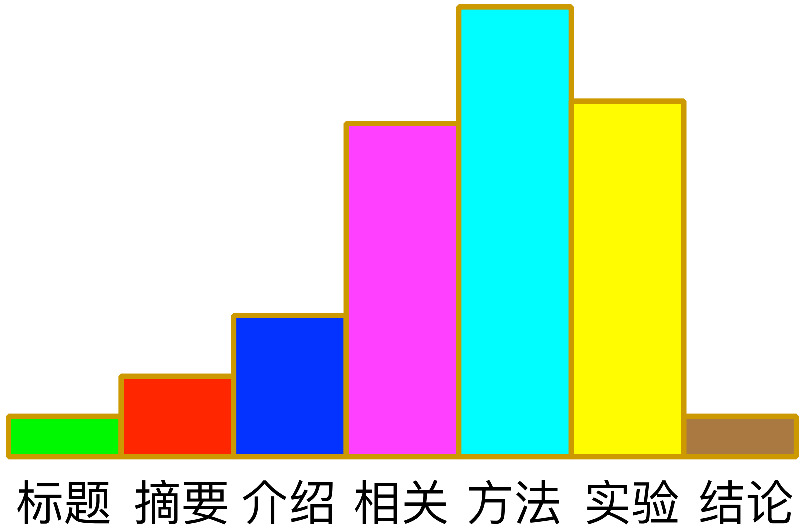
1 介绍
2 背景
3 方法
4 实验
5 相关工作
6 结论

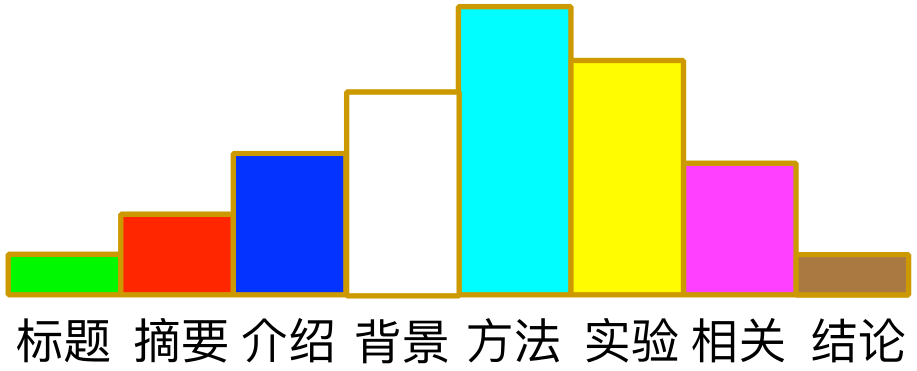


降低信息理解难度是关键

1 介绍
2 相关工作
3 方法
4 实验
5 结论

1 介绍
2 背景
3 方法
4 实验
5 相关工作
6 结论





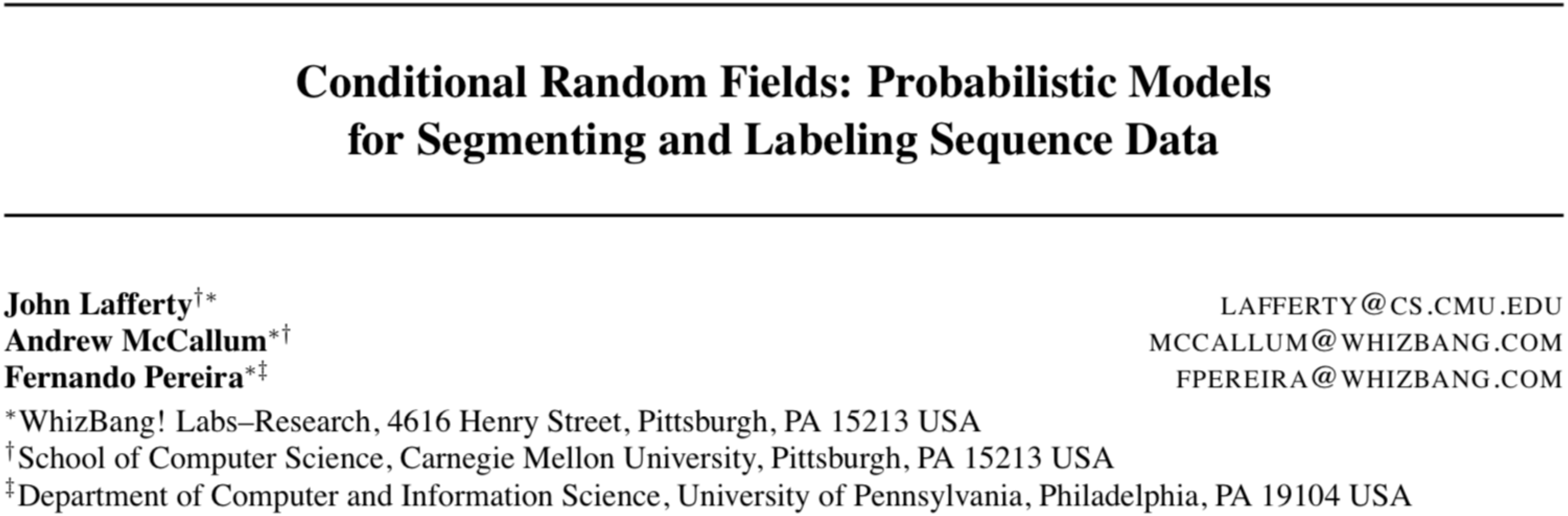
标题的写作技巧

标题的重要性

- 如何看浩如烟海的文献?
  - 根据标题过滤50%
  - 根据摘要再过滤20%
  - 根据介绍再过滤20%
  - 剩下的10%再仔细看论文

黄铠

例子



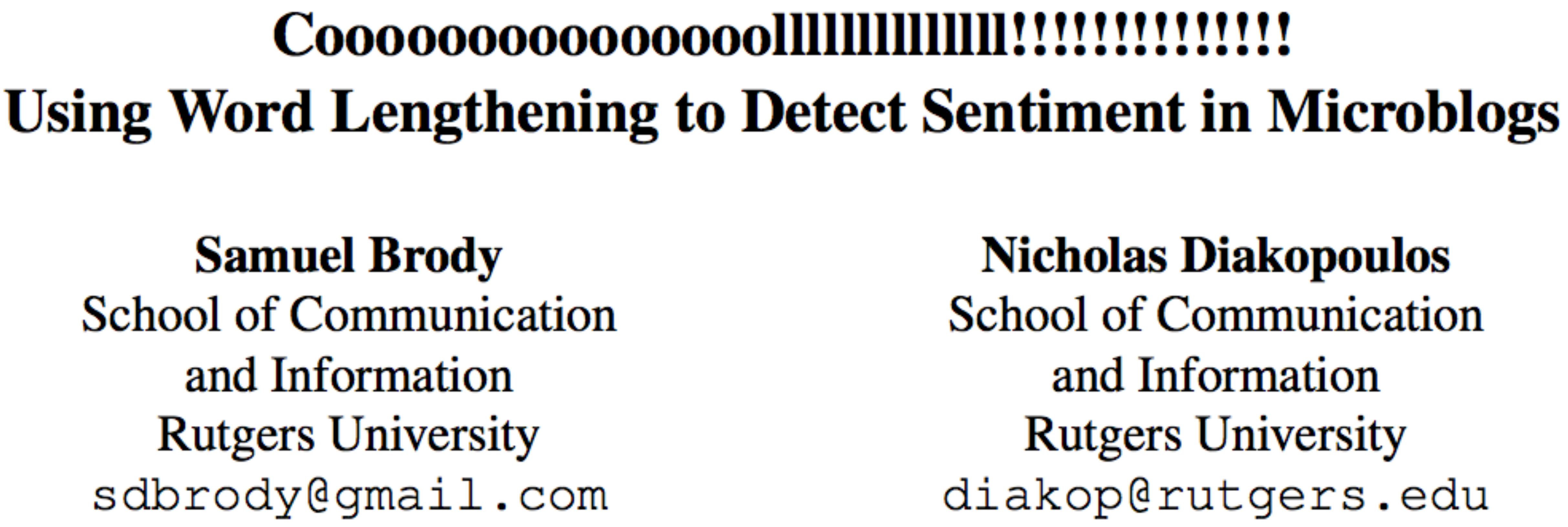
Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

John Lafferty[†,*] LAFFERTY@CS.CMU.EDU
Andrew McCallum[*,†] MCCALLUM@WHIZBANG.COM
Fernando Pereira[*,‡] FPEREIRA@WHIZBANG.COM

[*] WhizBang! Labs–Research, 4616 Henry Street, Pittsburgh, PA 15213 USA
[†] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA

- 用一句话概括你所做的工作
- 考虑搜索引擎的影响，包含关键词

例子



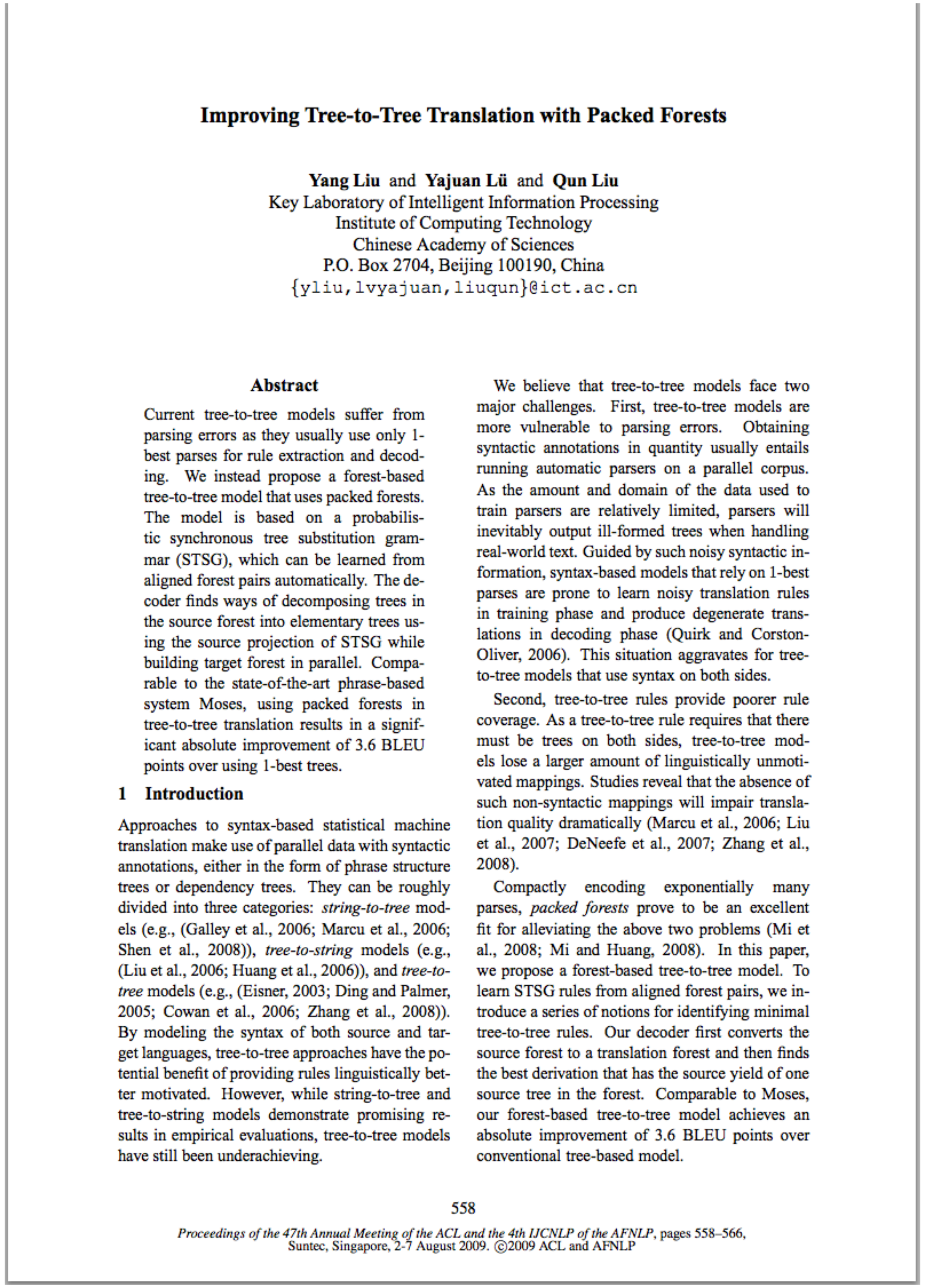
Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!
Using Word Lengthening to Detect Sentiment in Microblogs

Samuel Brody
School of Communication
and Information
Rutgers University
sdbrody@gmail.com

Nicholas Diakopoulos
School of Communication
and Information
Rutgers University
diakop@rutgers.edu

- 可以适当地别出心裁

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

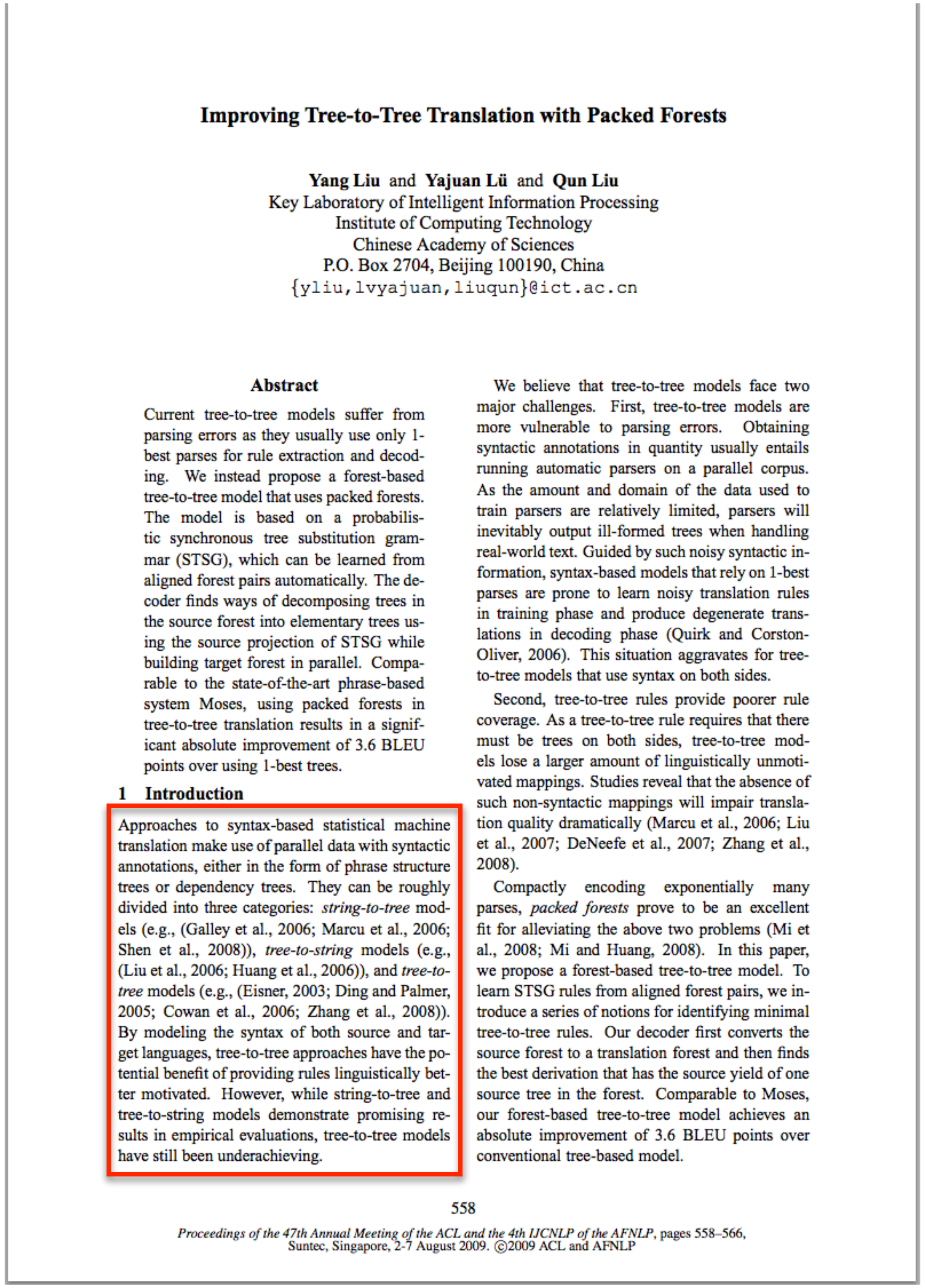
Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

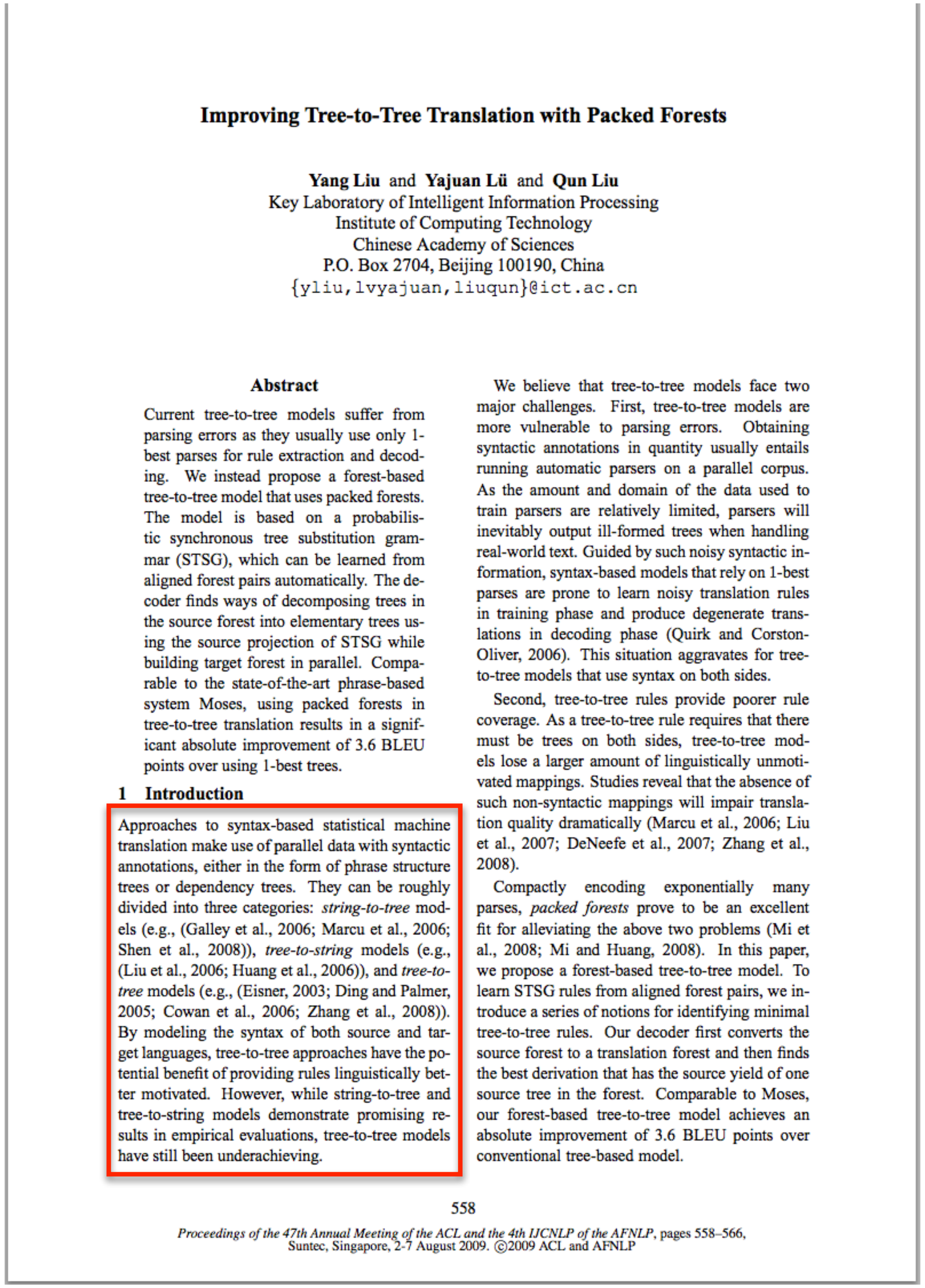
558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

问题

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

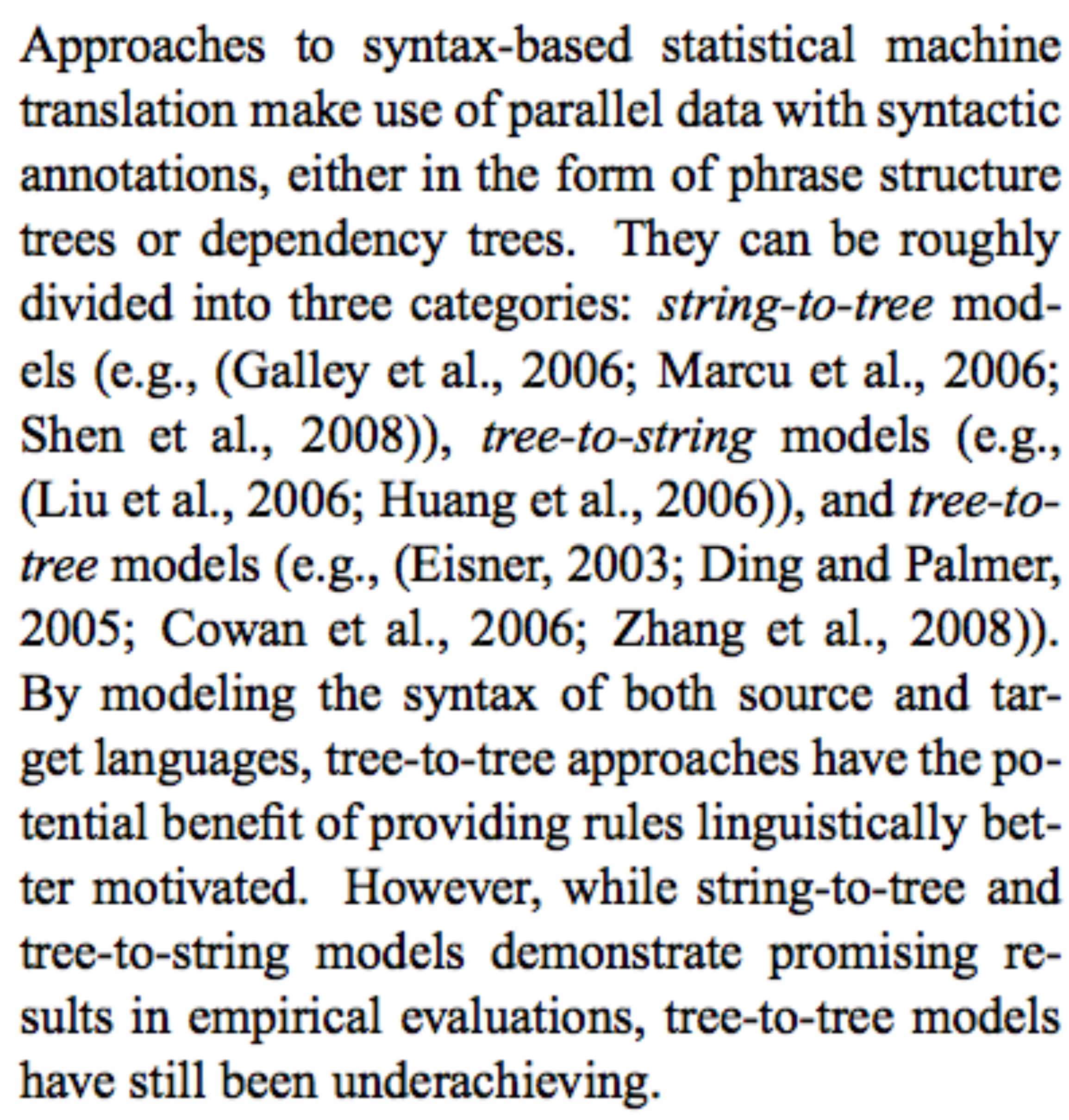
We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

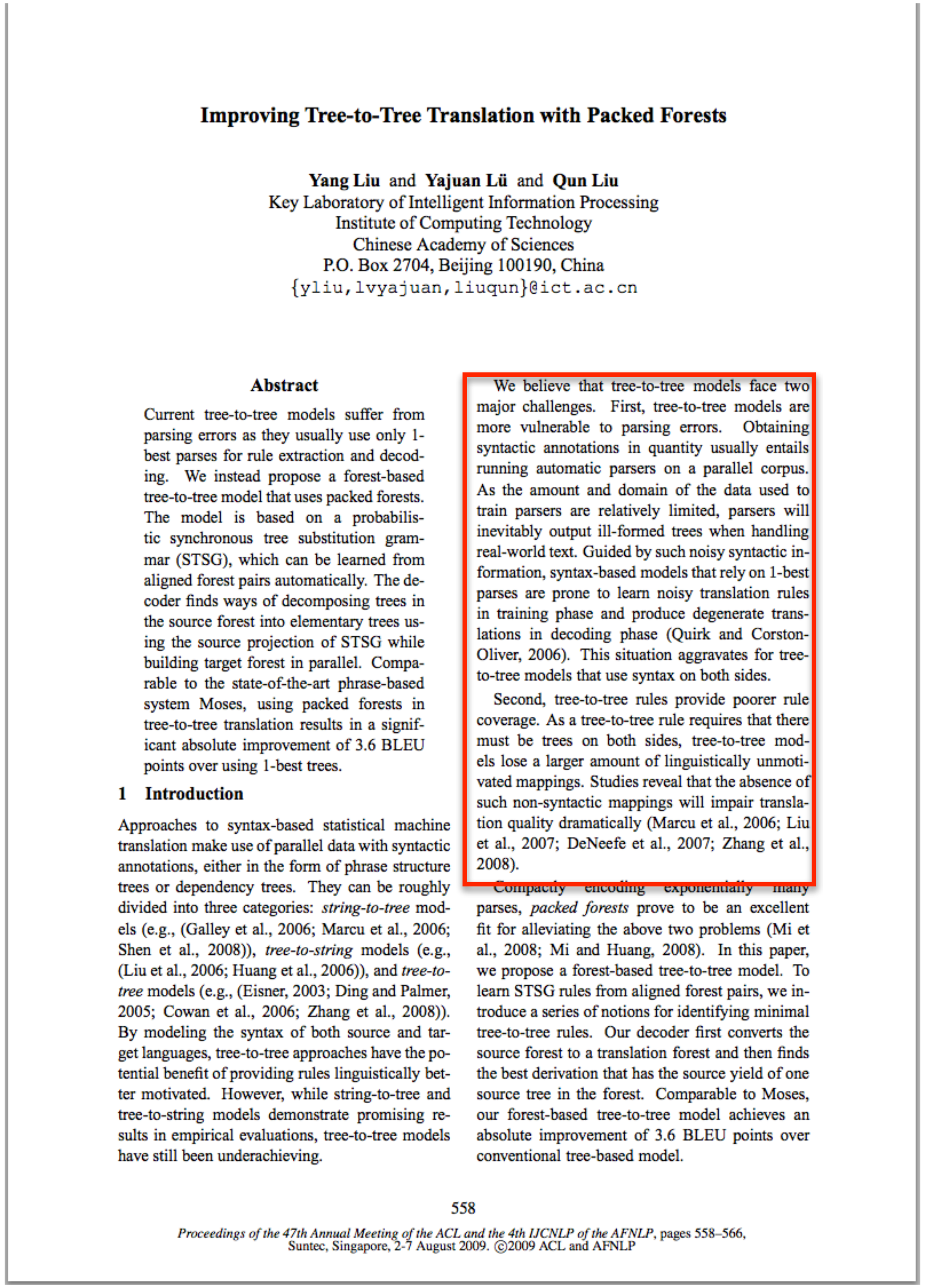
问题



Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

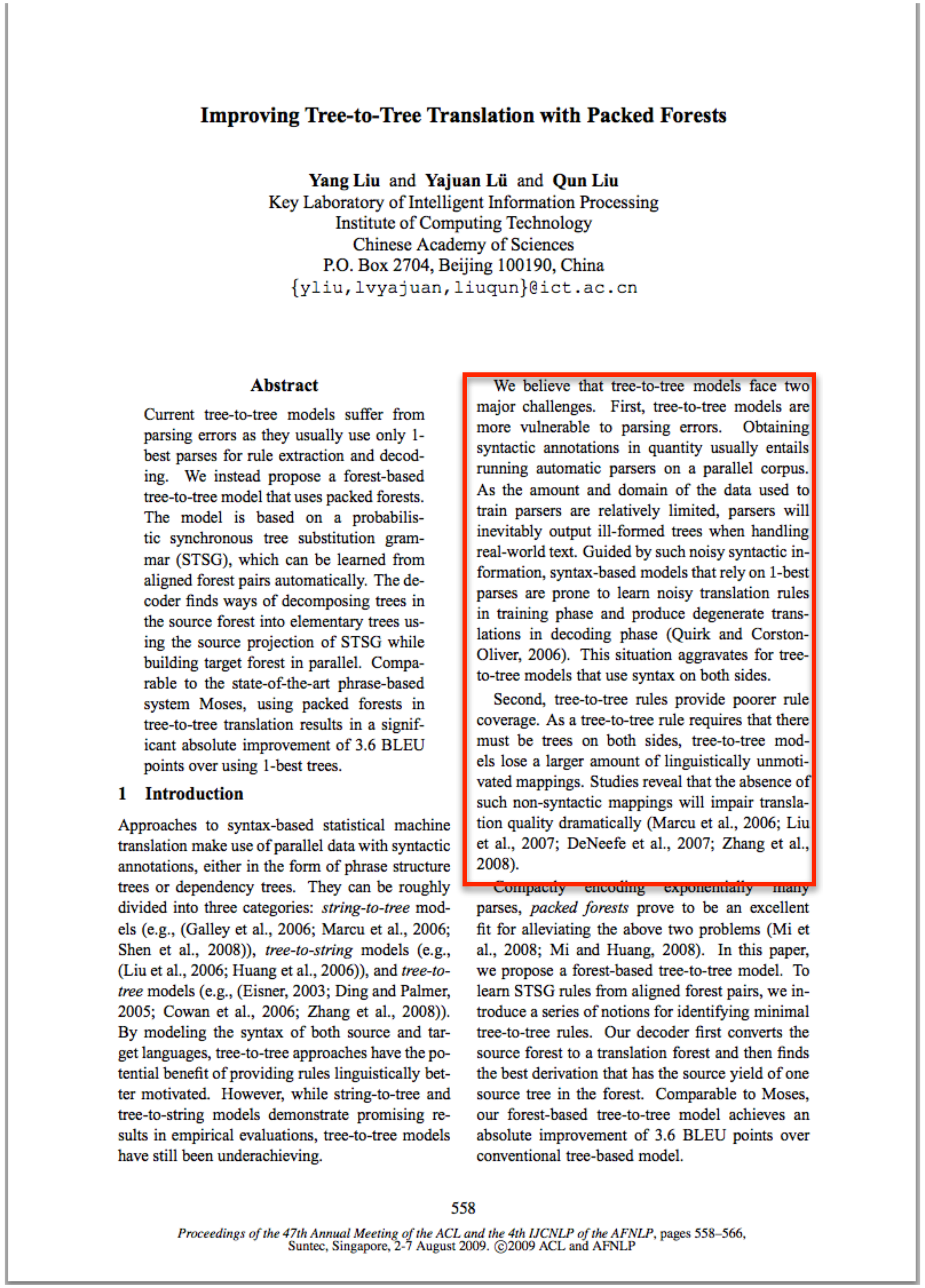
Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

挑战

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

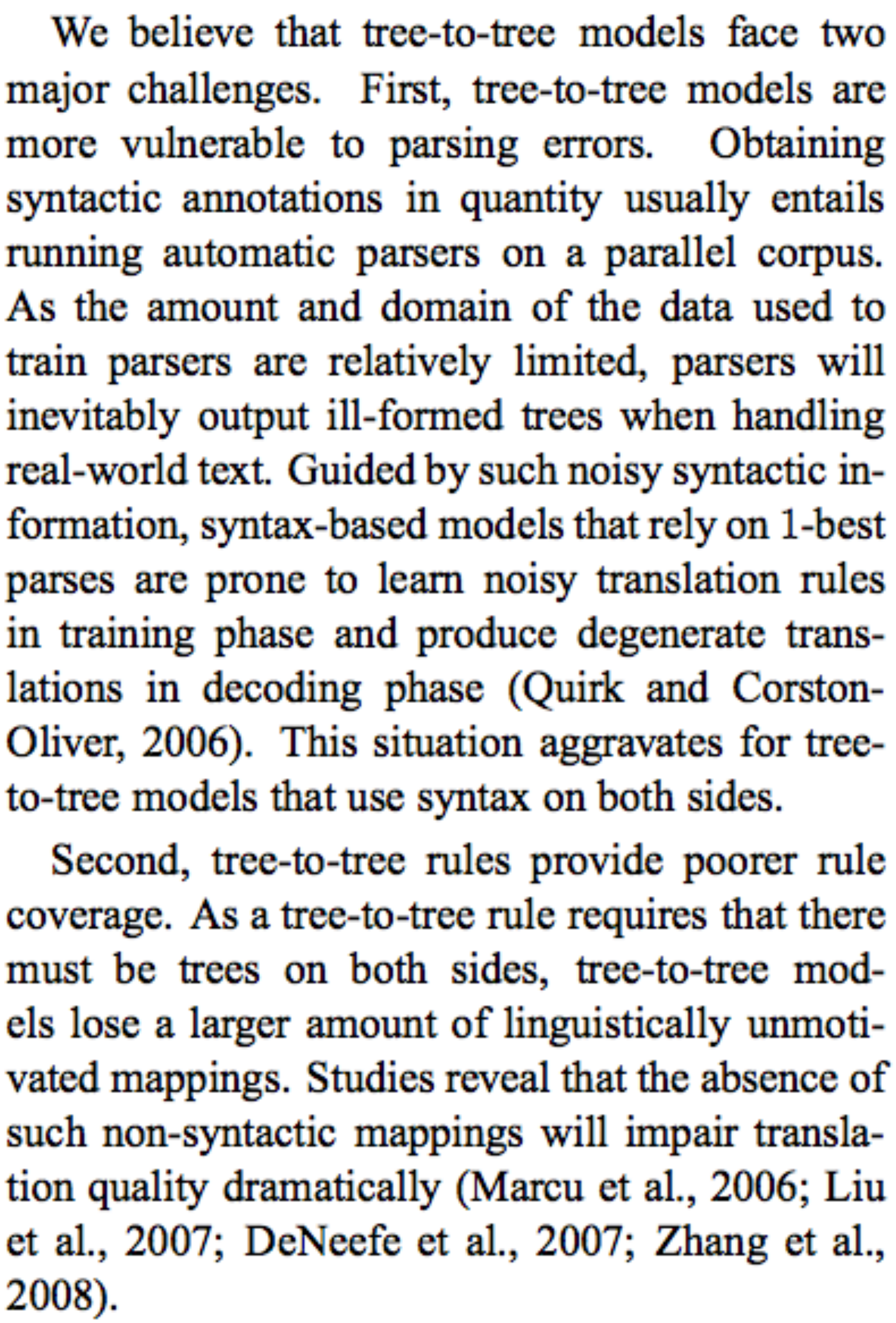
Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

挑战



We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

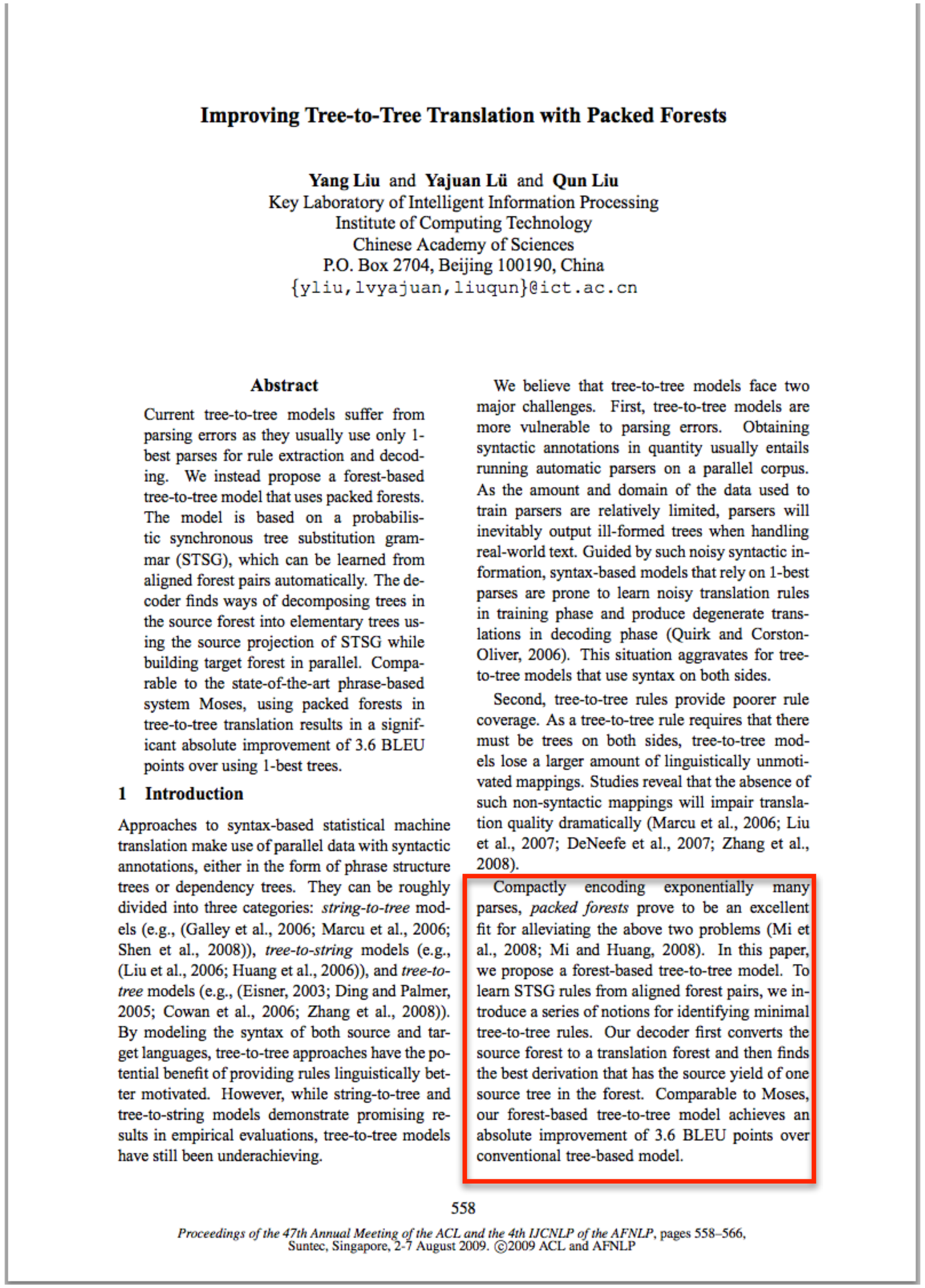
558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

我们的工作

Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

例子



Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

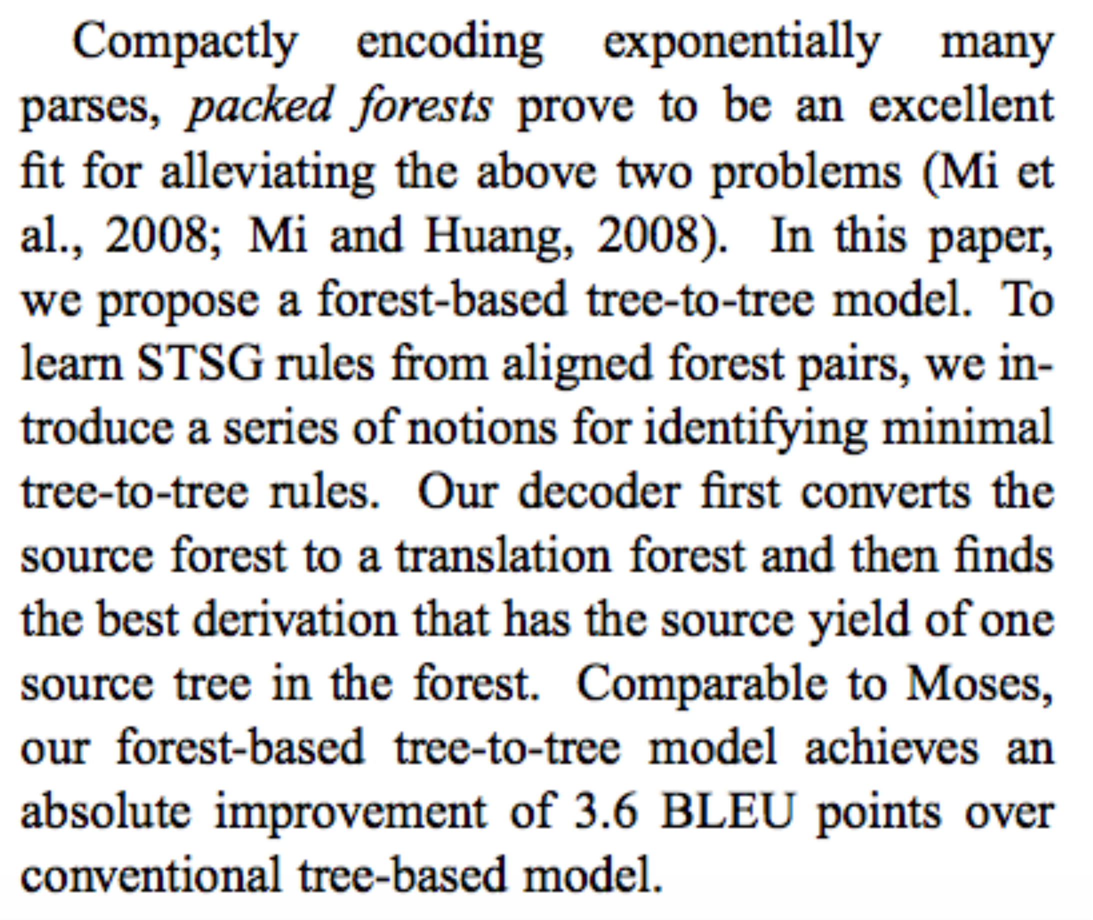
Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

558

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009. ©2009 ACL and AFNLP

我们的工作



Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

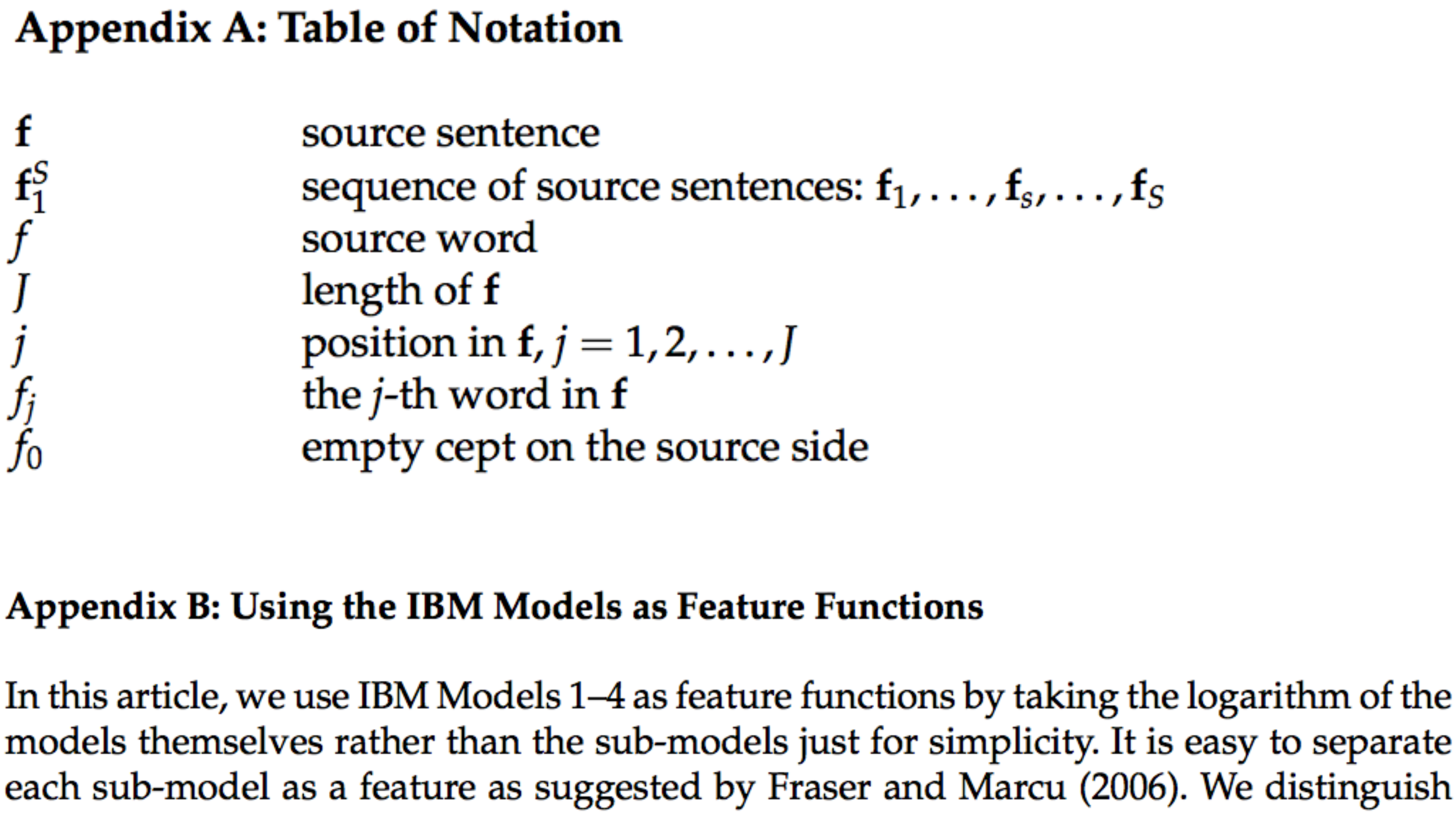
Yang Liu, Yajuan Lv, and Qun Liu. **Improving Tree-to-Tree Translation with Packed Forests**. In *ACL 2009*.

附录的写作技巧

附录

- 并非必需，但是对于读者深入理解你的工作有帮助，往往非常形式化
  - 证明
  - “鸡肋”
- 恰当地使用附录能显著提升论文的可读性

例子



Appendix A: Table of Notation

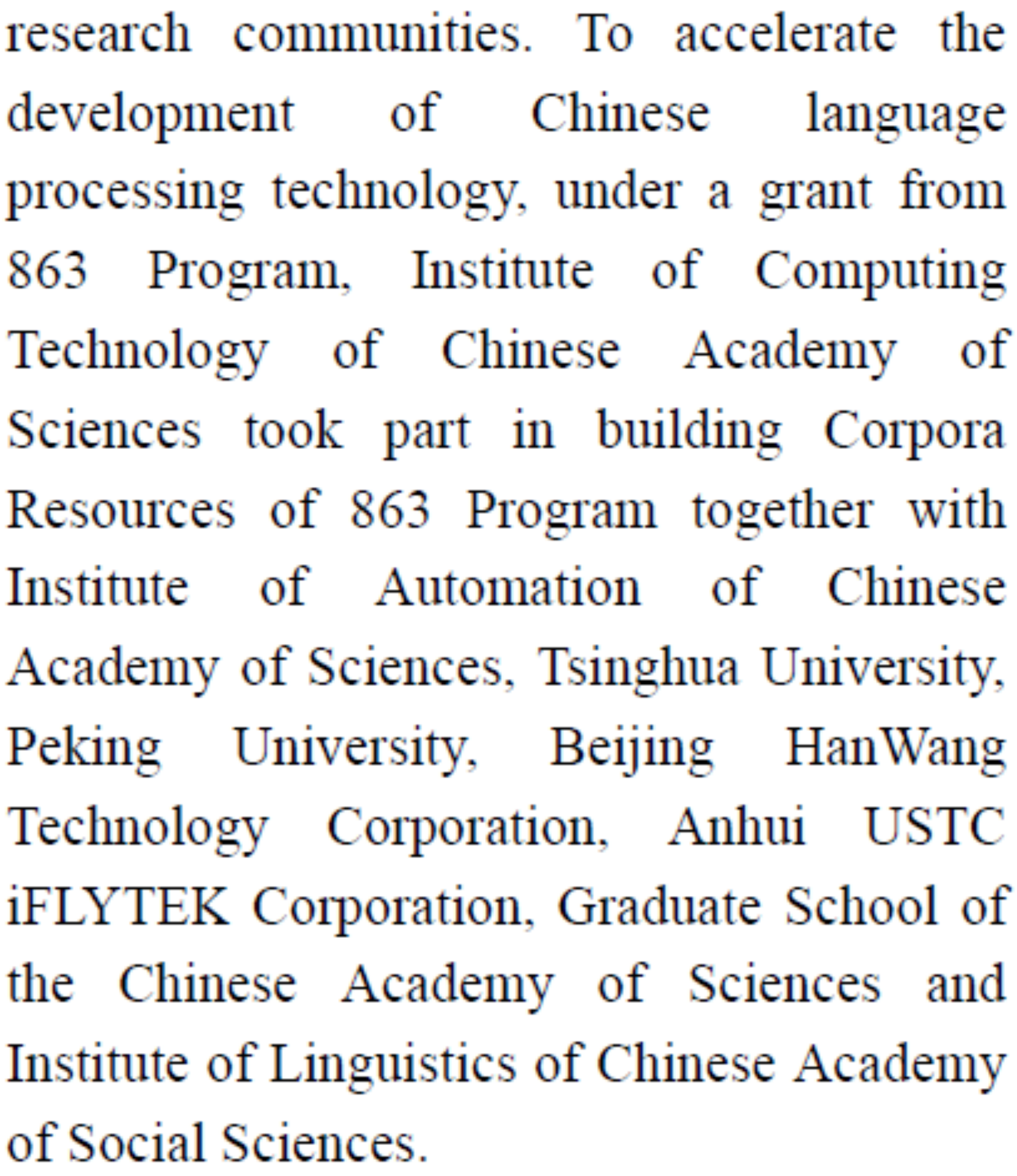
$\mathbf{f}$	source sentence
$\mathbf{f}_1^S$	sequence of source sentences: $\mathbf{f}_1, \ldots, \mathbf{f}_s, \ldots, \mathbf{f}_S$
f	source word
J	length of $\mathbf{f}$
j	position in $\mathbf{f}$, $j = 1, 2, \ldots, J$
f_j	the j-th word in $\mathbf{f}$
f_0	empty cept on the source side

Appendix B: Using the IBM Models as Feature Functions

In this article, we use IBM Models 1–4 as feature functions by taking the logarithm of the models themselves rather than the sub-models just for simplicity. It is easy to separate each sub-model as a feature as suggested by Fraser and Marcu (2006). We distinguish

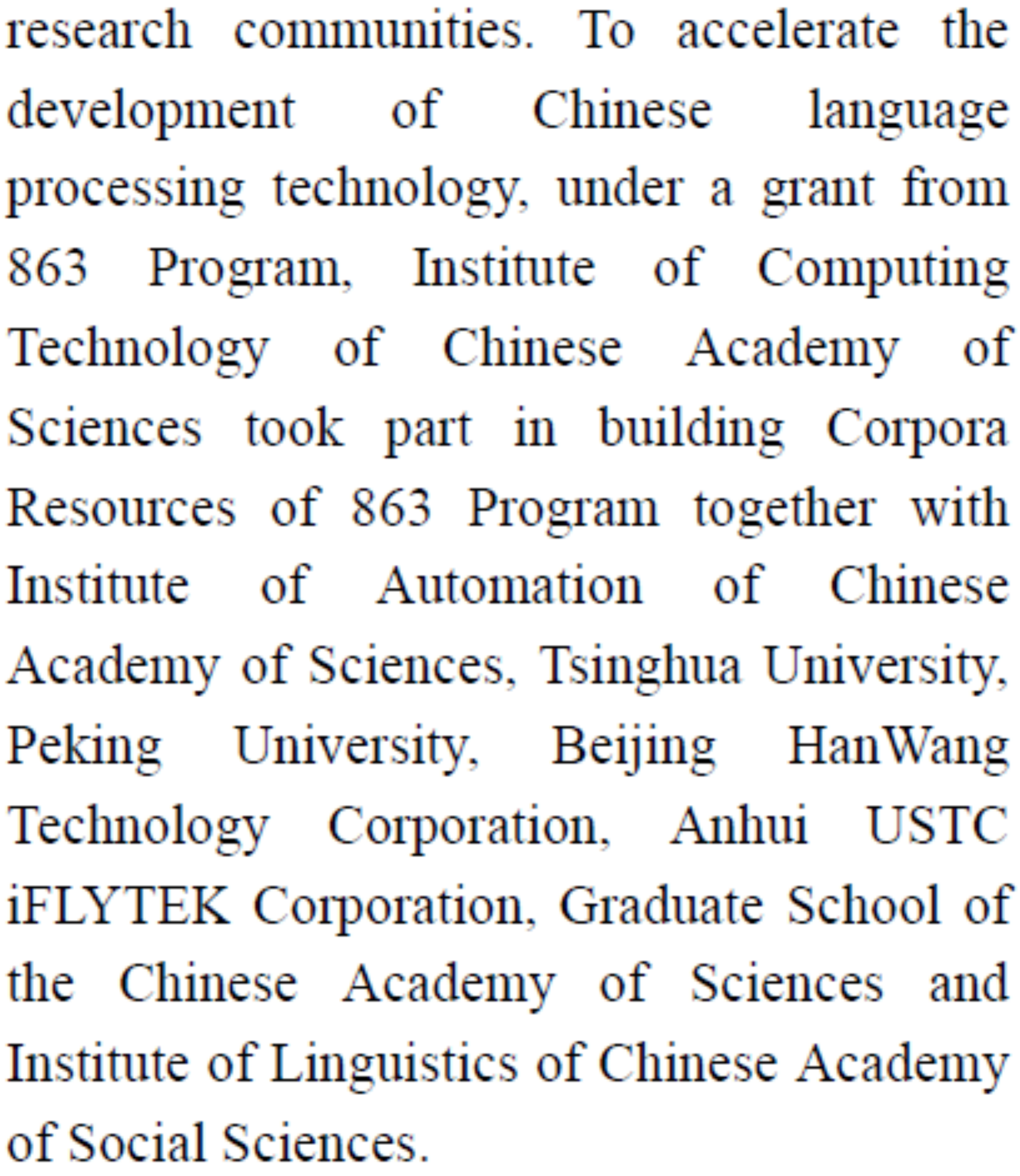
Yang Liu, Qun Liu, and Shouxun Lin. **Discriminative Word Alignment by Linear Modeling**. *Computational Linguistics*. 2010.

句子过长



research communities. To accelerate the development of Chinese language processing technology, under a grant from 863 Program, Institute of Computing Technology of Chinese Academy of Sciences took part in building Corpora Resources of 863 Program together with Institute of Automation of Chinese Academy of Sciences, Tsinghua University, Peking University, Beijing HanWang Technology Corporation, Anhui USTC iFLYTEK Corporation, Graduate School of the Chinese Academy of Sciences and Institute of Linguistics of Chinese Academy of Social Sciences.

句子过长



research communities. To accelerate the development of Chinese language processing technology, under a grant from 863 Program, Institute of Computing Technology of Chinese Academy of Sciences took part in building Corpora Resources of 863 Program together with Institute of Automation of Chinese Academy of Sciences, Tsinghua University, Peking University, Beijing HanWang Technology Corporation, Anhui USTC iFLYTEK Corporation, Graduate School of the Chinese Academy of Sciences and Institute of Linguistics of Chinese Academy of Social Sciences.

To advance the state of the art of Chinese language processing technology, many institutions in China took part in building the Corpora Resources under the grant from the 863 Program. These institutions include …

写作常见问题

- 经常使用被动句式
- 结构松散、口语化
- 不定冠词和定冠词的使用

被动句式+弱动词

The whole process of finding fuzzy-matched word pairs and computing their similarity is demonstrated in detail. More importantly, the performance of BLEU is significantly improved by integrating fuzzy matching.

被动句式+弱动词

The whole process of finding fuzzy-matched word pairs and computing their similarity is demonstrated in detail. More importantly, the performance of BLEU is significantly improved by integrating fuzzy matching.

被动句式+弱动词

The whole process of finding fuzzy-matched word pairs and computing their similarity is demonstrated in detail. More importantly, the performance of BLEU is significantly improved by integrating fuzzy matching.

⇩

We demonstrate how to find fuzzy-matched word pairs and compute their similarities in detail. More importantly, integrating fuzzy matching significantly improved the translation performance in terms of BLEU.

结构松散+口语化+缺乏力度

In this step, we want to induce an alignment between words and predicates. The alignment can give a roughly mapping between words and the predicates that express their meanings, so it would be a useful constraint for rule extraction and reduce the searching space.

结构松散+口语化+缺乏力度

In this step, we want to induce an alignment between words and predicates. The alignment can give a roughly mapping between words and the predicates that express their meanings, so it would be a useful constraint for rule extraction and reduce the searching space.

结构松散+口语化+缺乏力度

In this step, we want to induce an alignment between words and predicates. The alignment can give a roughly mapping between words and the predicates that express their meanings, so it would be a useful constraint for rule extraction and reduce the searching space.

This step induces an alignment between words and predicates. Reflecting a rough mapping between natural languages and logic, such alignments impose linguistically motivated constraints on the search space and improve the efficiency of rule extraction.

a还是an

A *FBI agent* or ***An*** *FBI agent?*

A *FIFA officer* or ***An*** *FIFA officer?*

a还是an

An *FBI agent?*

A *FIFA officer* or ***An*** *FIFA officer?*

a还是an

An *FBI agent?*

A *FIFA officer*

a还是an

***An** FBI agent?*

***A** FIFA officer*

看发音而不是字母

a还是an

An *FBI agent?*

A *FIFA officer*

看发音而不是字母

SVM F-score X-ray

NBA CRF European

怎么使用the?

The statistical translation models that try to capture the recursive structures of the language over the last several years.

In the experiments on the Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

怎么使用the?

~~The~~ statistical translation models that try to capture the recursive structures of the language over the last several years.

In the experiments on the Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

怎么使用the?

~~The~~ statistical translation models that try to capture the recursive structures of ~~the~~ language over the last several years.

In the experiments on the Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

怎么使用the?

~~The~~ statistical translation models that try to capture the recursive structures of ~~the~~ language over the last several years.

In ~~the~~ experiments on the Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

怎么使用the?

~~The~~ statistical translation models that try to capture the recursive structures of ~~the~~ language over the last several years.

In ~~the~~ experiments on ~~the~~ Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

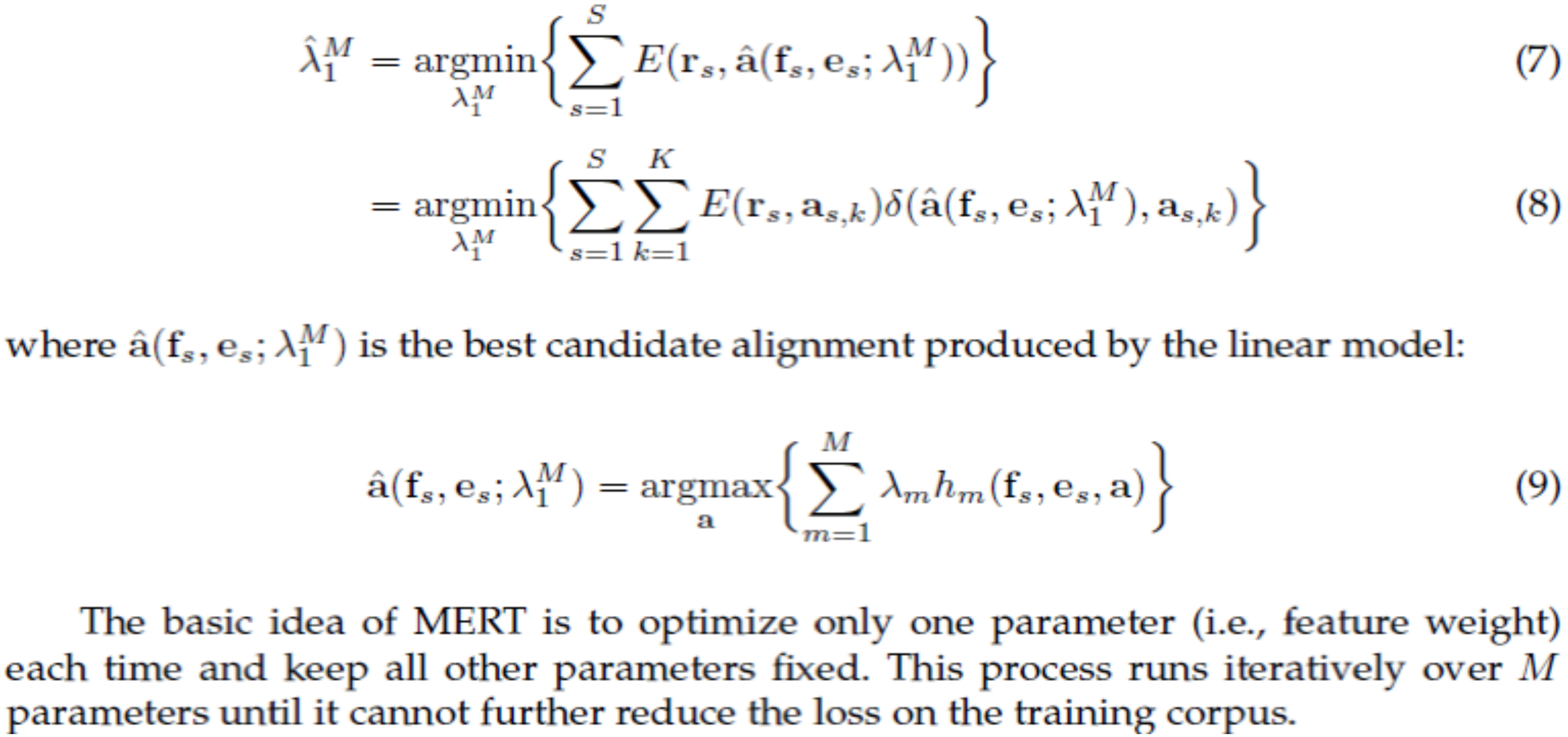
怎么使用the?

~~The~~ statistical translation models that try to capture the recursive structures of ~~the~~ language over the last several years.

In ~~the~~ experiments on ~~the~~ Chinese-English translation, we find that the model chooses to build ~~the~~ structures that are more syntactic.

“the”一般指特指，否则不加

公式的缩进



$$\hat{\lambda}_1^M = \operatorname*{argmin}_{\lambda_1^M}\left\{\sum_{s=1}^{S} E(\mathbf{r}_s, \hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M))\right\} \tag{7}$$

$$= \operatorname*{argmin}_{\lambda_1^M}\left\{\sum_{s=1}^{S}\sum_{k=1}^{K} E(\mathbf{r}_s, \mathbf{a}_{s,k})\delta(\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M), \mathbf{a}_{s,k})\right\} \tag{8}$$

where $\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M)$ is the best candidate alignment produced by the linear model:

$$\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M) = \operatorname*{argmax}_{\mathbf{a}}\left\{\sum_{m=1}^{M} \lambda_m h_m(\mathbf{f}_s, \mathbf{e}_s, \mathbf{a})\right\} \tag{9}$$

The basic idea of MERT is to optimize only one parameter (i.e., feature weight) each time and keep all other parameters fixed. This process runs iteratively over M parameters until it cannot further reduce the loss on the training corpus.

当公式后的文本与公式有关，则不缩进，否则缩进

引用的写法

Jack (2010) argues that it is important to use syntax.

This algorithm proves to runs in approximately linear time (Jack, 2010).

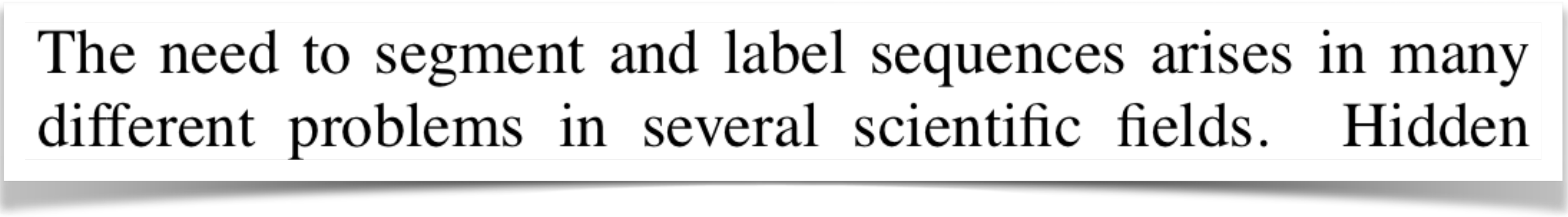
前者表示人，后者去掉应该不影响整句话的意思。

如何提高英语写作?

提高英语写作的窍门

- 找著名学者（尤其是native speaker）的论文钻研，学习句式和词汇用法，做笔记
- 写作时手边放一部纸质词典，经常翻看
- 拿不准的地方找Google：双引号查询

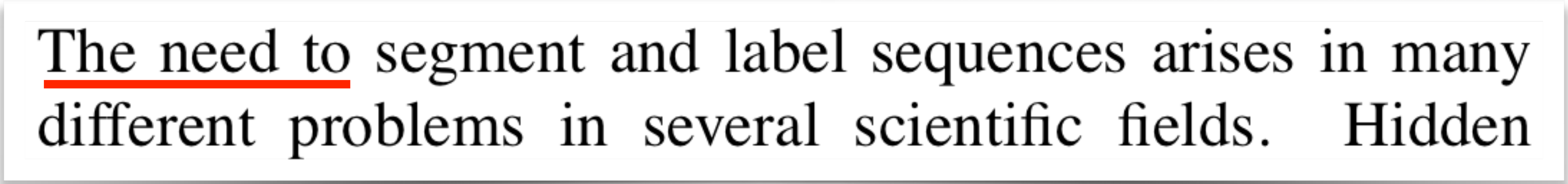
学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

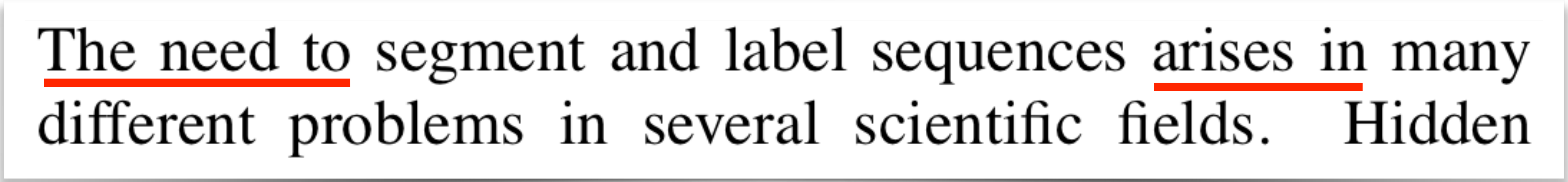
学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

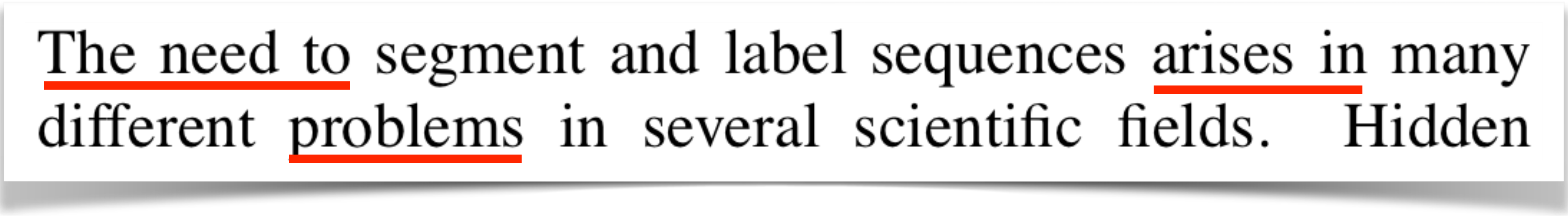
学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

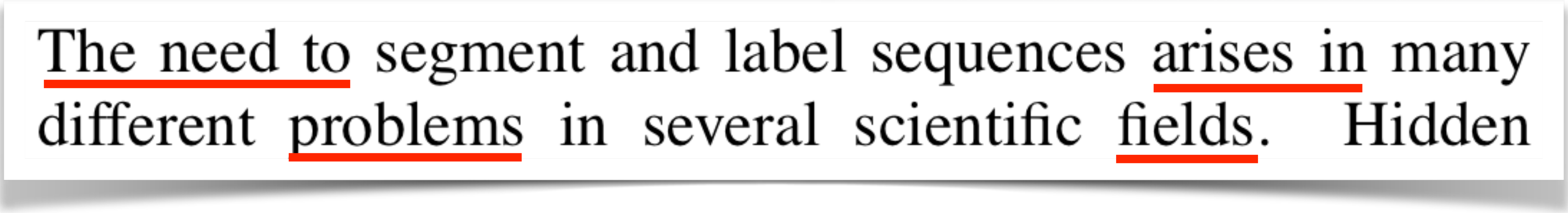
学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

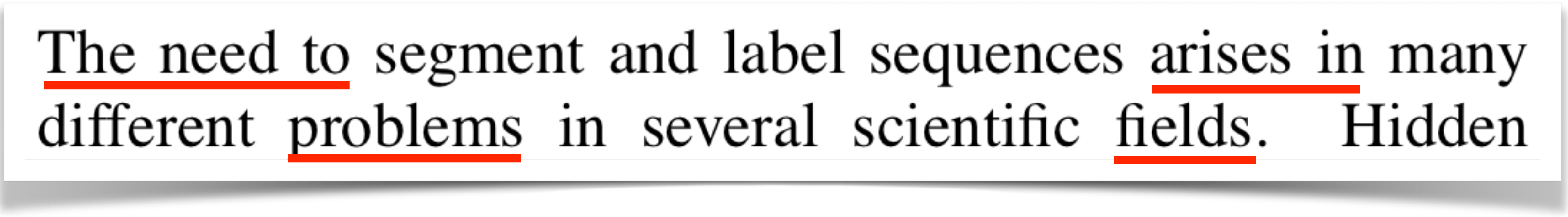
学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

学习句式和用法

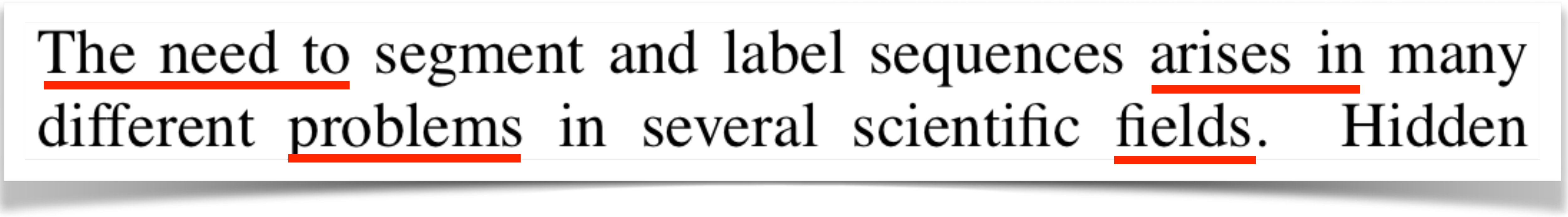


The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

学习句式和用法

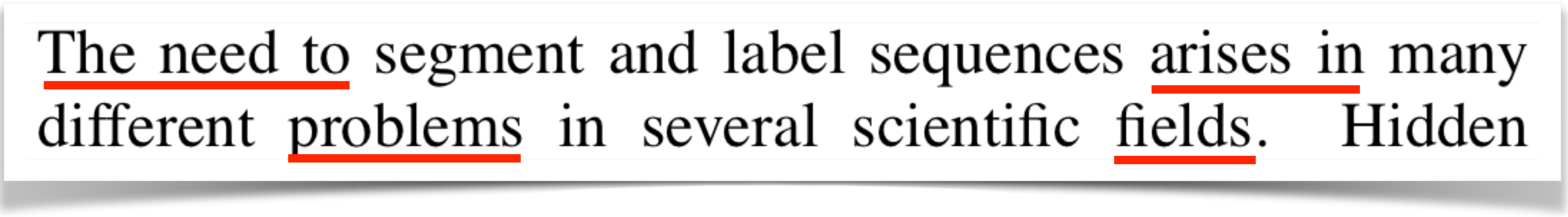


The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式 *the need to … arises in … problems (fields)*

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

学习句式和用法



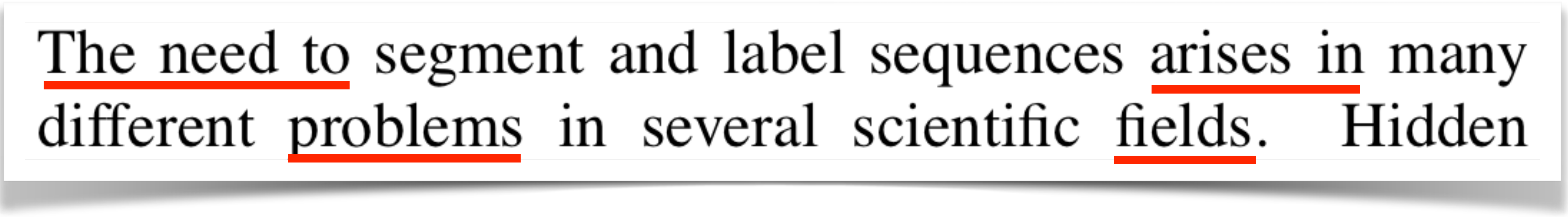
The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式 *the need to … arises in … problems (fields)*

造句

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

学习句式和用法



The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式 *the need to … arises in … problems (fields)*

造句

The need to learn latent-variable models from unlabeled data arises in many NLP problems.

John Lafferty, Andrew McCallum, and Fernando Pereira. **Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data**. In *ICML 2003*.

利用搜索引擎

Maximizing the likelihood ______ the training data.

(A) in (B) on (c) of

利用搜索引擎

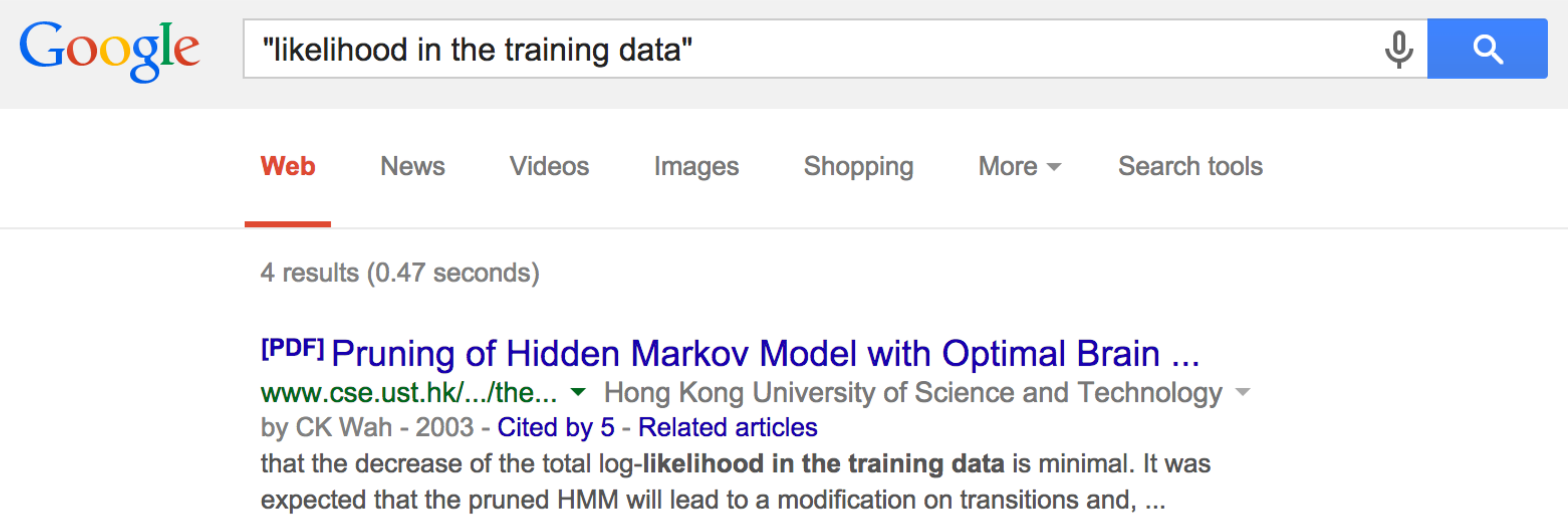
Maximizing the likelihood _______ the training data.

(A) in (B) on (c) of

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in   (B) on   (c) of



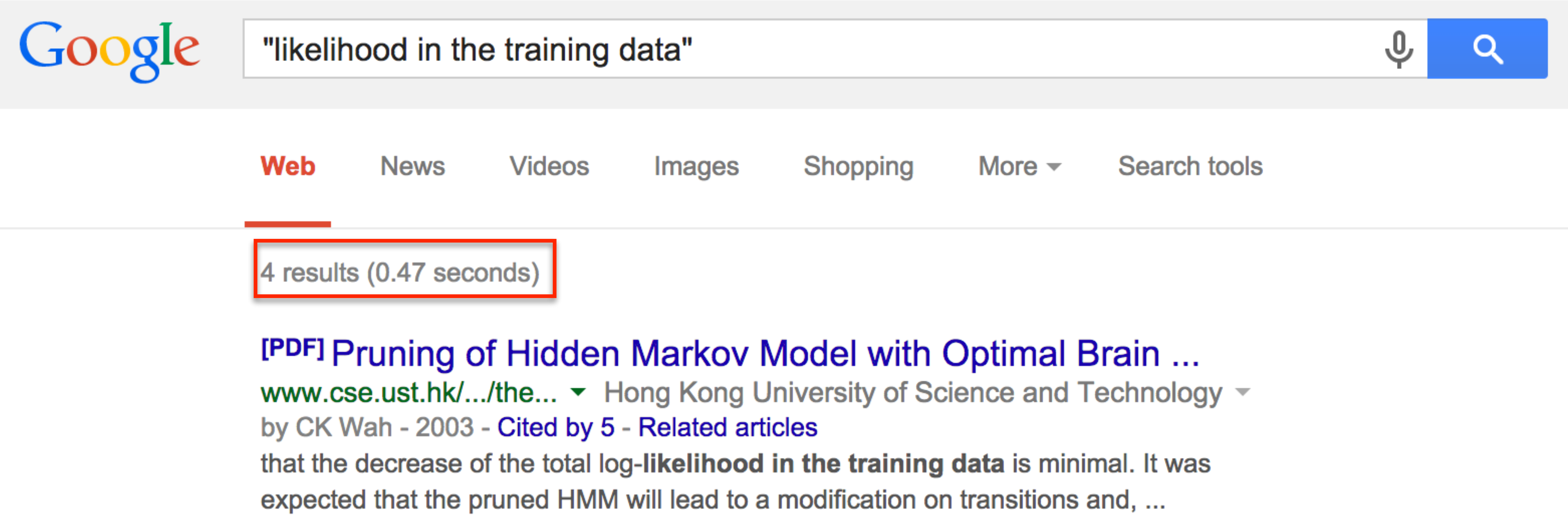
4 results (0.47 seconds)

[PDF] Pruning of Hidden Markov Model with Optimal Brain ...
www.cse.ust.hk/.../the... Hong Kong University of Science and Technology
by CK Wah - 2003 - Cited by 5 - Related articles
that the decrease of the total log-**likelihood in the training data** is minimal. It was expected that the pruned HMM will lead to a modification on transitions and, ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of



"likelihood in the training data"

4 results (0.47 seconds)

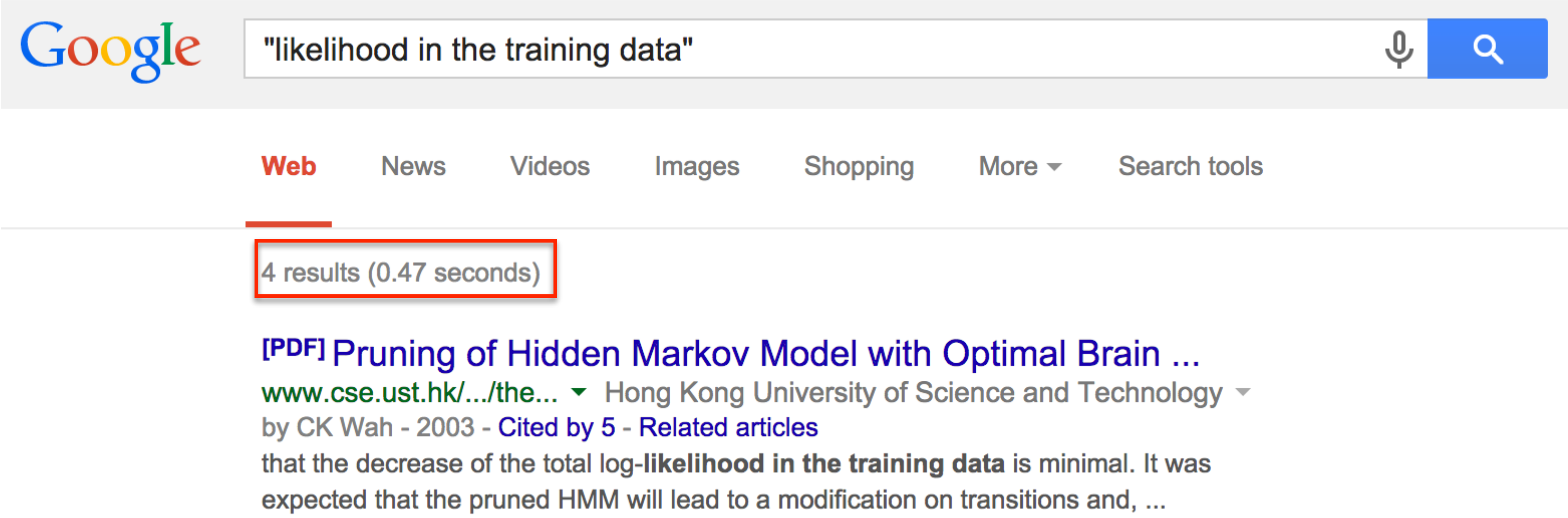
[PDF] Pruning of Hidden Markov Model with Optimal Brain ...
www.cse.ust.hk/.../the... Hong Kong University of Science and Technology
by CK Wah - 2003 - Cited by 5 - Related articles
that the decrease of the total log-**likelihood in the training data** is minimal. It was expected that the pruned HMM will lead to a modification on transitions and, ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4



"likelihood in the training data"

4 results (0.47 seconds)

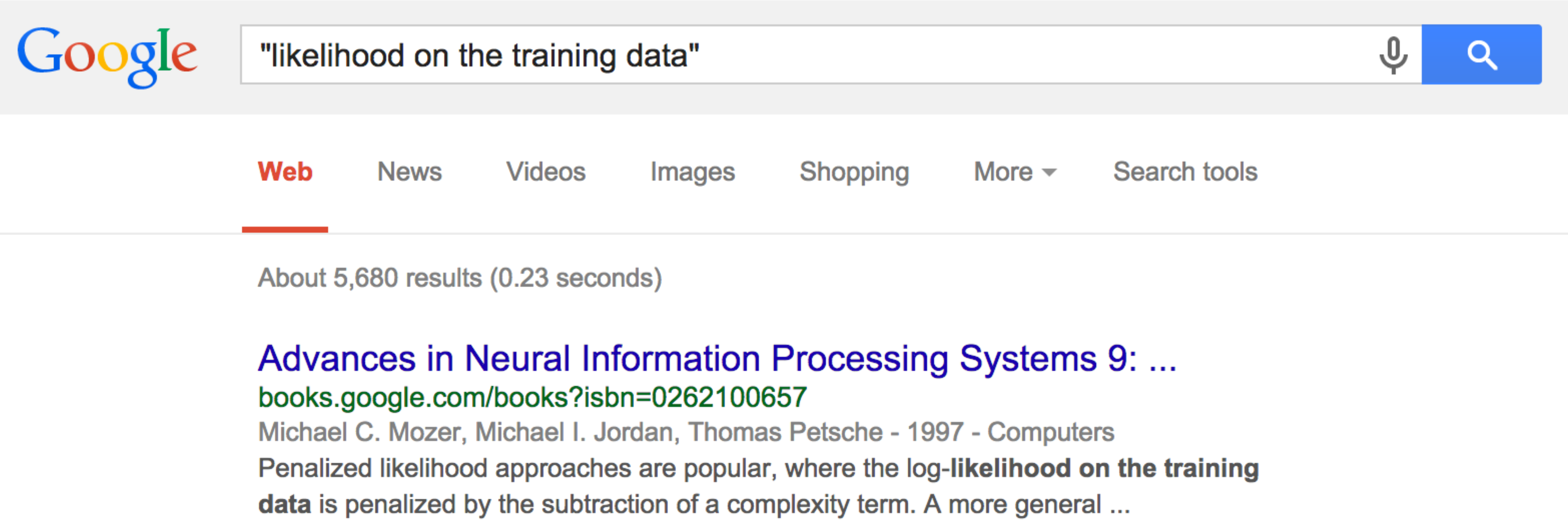
[PDF] Pruning of Hidden Markov Model with Optimal Brain ...
www.cse.ust.hk/.../the... Hong Kong University of Science and Technology
by CK Wah - 2003 - Cited by 5 - Related articles
that the decrease of the total log-**likelihood in the training data** is minimal. It was expected that the pruned HMM will lead to a modification on transitions and, ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4



"likelihood on the training data"

About 5,680 results (0.23 seconds)

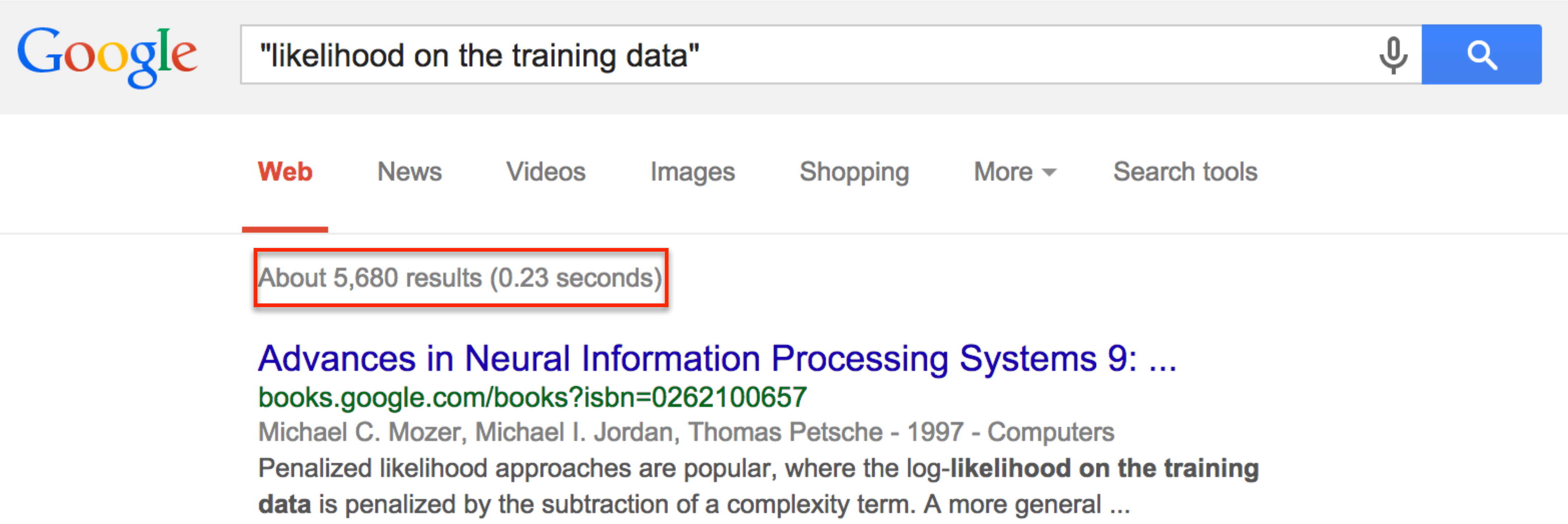
Advances in Neural Information Processing Systems 9: ...
books.google.com/books?isbn=0262100657
Michael C. Mozer, Michael I. Jordan, Thomas Petsche - 1997 - Computers
Penalized likelihood approaches are popular, where the log-**likelihood on the training data** is penalized by the subtraction of a complexity term. A more general ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4



"likelihood on the training data"

About 5,680 results (0.23 seconds)

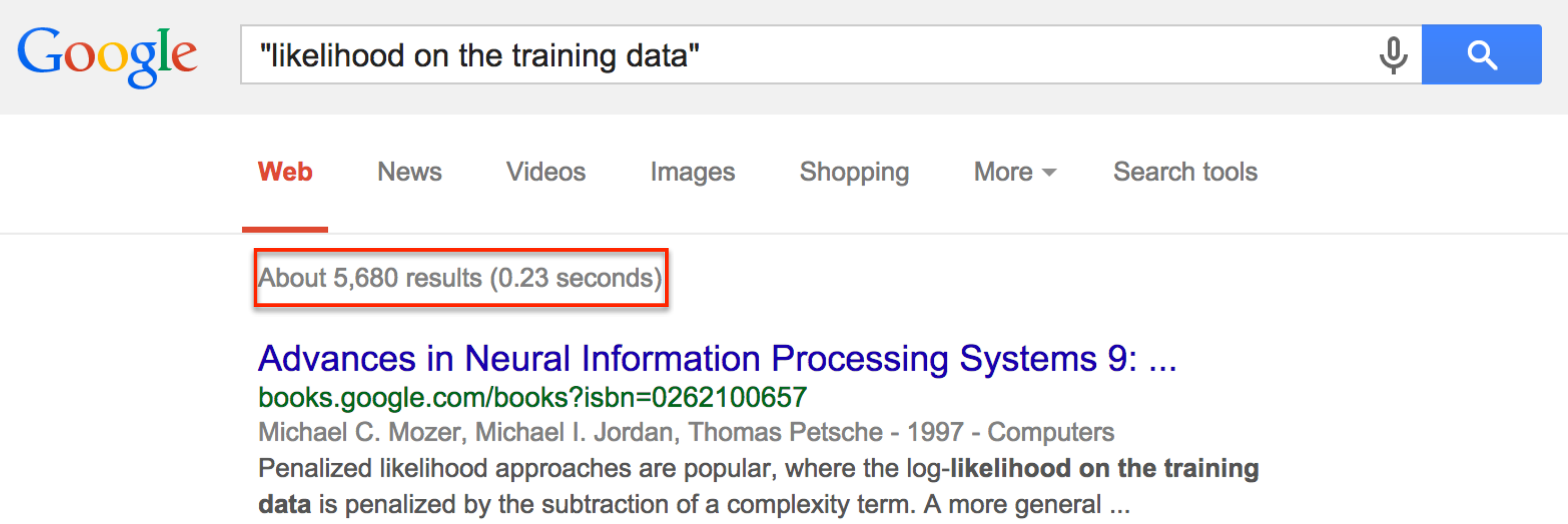
Advances in Neural Information Processing Systems 9: ...
books.google.com/books?isbn=0262100657
Michael C. Mozer, Michael I. Jordan, Thomas Petsche - 1997 - Computers
Penalized likelihood approaches are popular, where the log-**likelihood on the training data** is penalized by the subtraction of a complexity term. A more general ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4 5,680



"likelihood on the training data"

About 5,680 results (0.23 seconds)

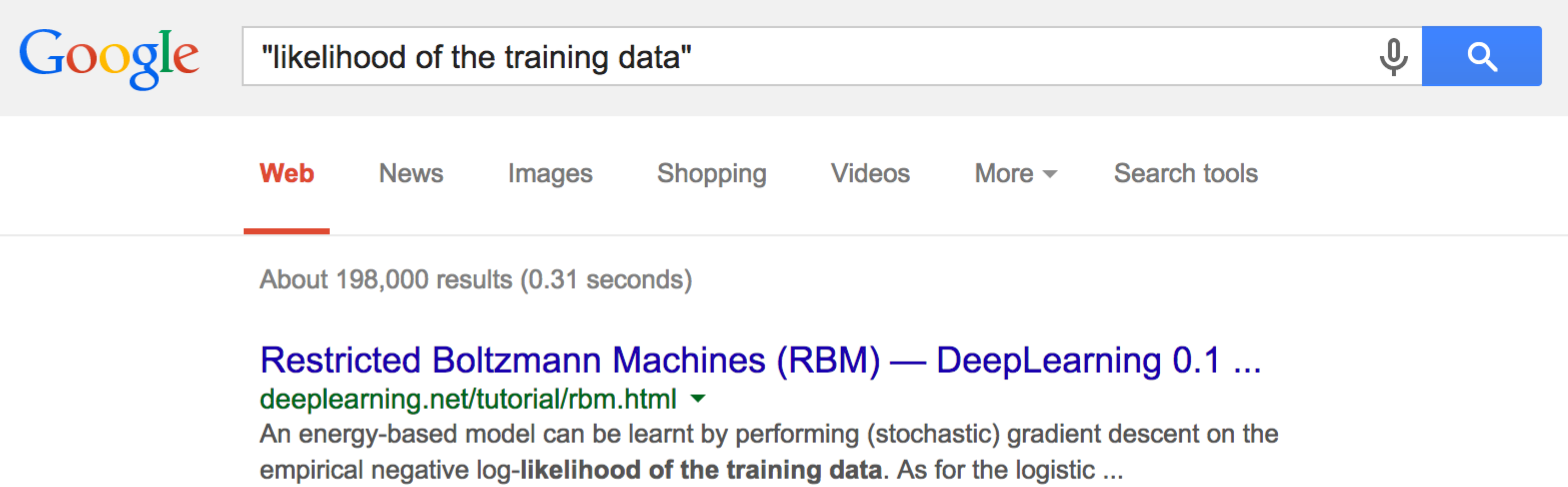
Advances in Neural Information Processing Systems 9: ...
books.google.com/books?isbn=0262100657
Michael C. Mozer, Michael I. Jordan, Thomas Petsche - 1997 - Computers
Penalized likelihood approaches are popular, where the log-**likelihood on the training data** is penalized by the subtraction of a complexity term. A more general ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4 5,680



About 198,000 results (0.31 seconds)

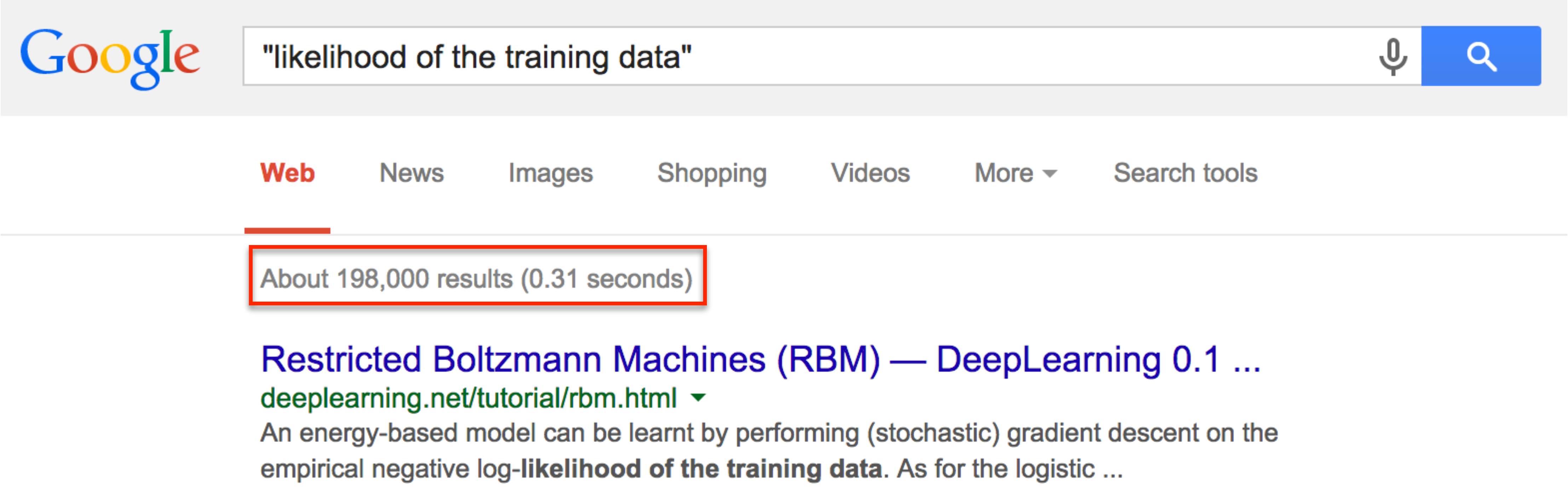
Restricted Boltzmann Machines (RBM) — DeepLearning 0.1 ...
deeplearning.net/tutorial/rbm.html
An energy-based model can be learnt by performing (stochastic) gradient descent on the empirical negative log-**likelihood of the training data**. As for the logistic ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4 5,680



"likelihood of the training data"

About 198,000 results (0.31 seconds)

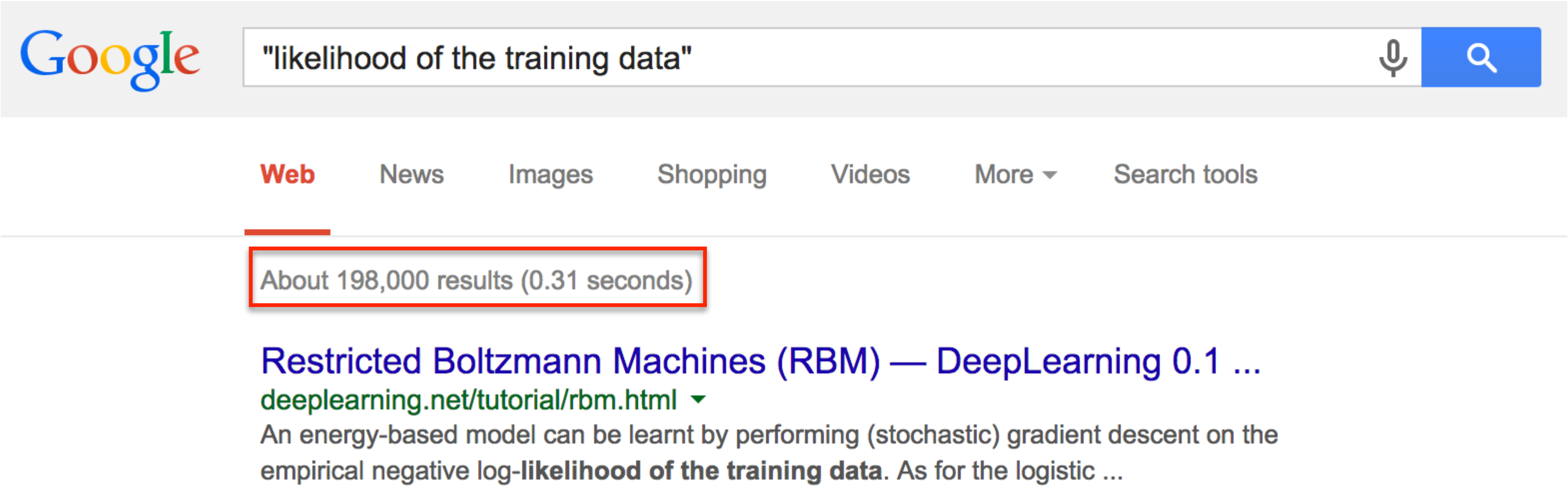
Restricted Boltzmann Machines (RBM) — DeepLearning 0.1 ...
deeplearning.net/tutorial/rbm.html
An energy-based model can be learnt by performing (stochastic) gradient descent on the empirical negative log-**likelihood of the training data**. As for the logistic ...

利用搜索引擎

Maximizing the likelihood ________ the training data.

(A) in (B) on (c) of

4 5,680 198,000



"likelihood of the training data"

About 198,000 results (0.31 seconds)

Restricted Boltzmann Machines (RBM) — DeepLearning 0.1 ...
deeplearning.net/tutorial/rbm.html
An energy-based model can be learnt by performing (stochastic) gradient descent on the empirical negative log-**likelihood of the training data**. As for the logistic ...