

# A Hybrid Approach to Skeleton-based Translation

Tong Xiao, Jingbo Zhu, Chunliang Zhang

# 紹介文献

author = {Xiao, Tong and Zhu, Jingbo and Zhang, Chunliang},  
title = {A Hybrid Approach to Skeleton-based Translation},  
booktitle = {Proceedings of the 52nd Annual Meeting of the Association  
for Computational Linguistics (Volume 2: Short Papers)},  
month = {June},  
year = {2014},  
address = {Baltimore, Maryland},  
publisher = {Association for Computational Linguistics},  
pages = {563--568},  
url = {<http://www.aclweb.org/anthology/P14-2092>}

# はじめに

現存のSMTは入力文を単語単位ですべて訳しようとする

これでは、

- 対象単語それぞれが持つ役割は無視される。
- 手法も実際人間翻訳とは大きく異なる

直感的に人手翻訳はキー単語/文構造を翻訳し、残りを穴埋め

# モデル概要

- スケルトン・ベース・モデル
  - スケルトン・翻訳・モデル
  - フル・翻訳・モデル

# スケルトン・ベース・モデル

- スケルトンとは文中の重要な要素/構造を指す
- スケルトン抽出は幾通りか可能
  1. フル・シンタクティック・ツリーより部分木削除
  2. 非重要単語の削減かつ文法構造の保持
    - などなど
- 本稿では 2 番めの手法を適用

# スケルトン抽出・具体例

- The cost of seawater desalination treatment has been further reduced from 5 yuan per ton.



- The cost has been further reduced.

シンプリフィケーション問題と類似

# 動機

1. スケルトン文は文として整形されているため、SMTだけに限らず、様々なMTシステムで有効
2. 単語削減によるシンプリフィケーションの研究は多く研究されている

# モデル説明

$$\hat{t} = \arg \max_t P(t|\tau, s)$$

t: 目的語文

$\tau$ : スケルトン

s: 原語文



# スケルトン in SMT

SMTとはフレーズ・ペアと翻訳文法の導出の最適化

$$\hat{d} = \arg \max_d g(d)$$

d: 導出

g(): スコア関数

スケルトン・モデルでは

$$g(d) = g_{skel}(d) + g_{full}(d)$$

# スケルトン・モデル in SMT 例

phrase 1	每吨 海水淡化 处理 的 成本 在 5 元 的 基础 上 进 一 步 下 降 。
	<p>the cost <math>p_1</math></p> <p><b>Skeleton:</b> <math>g(d_\tau; \mathbf{w}^\tau, m, lm^\tau) = \mathbf{w}_m^\tau \cdot \mathbf{f}_m(p_1) + w_{lm}^\tau \cdot lm^\tau</math> ("the cost")</p> <p><b>Full:</b> <math>g(d; \mathbf{w}, m, lm) = \mathbf{w}_m \cdot \mathbf{f}_m(p_1) + w_{lm} \cdot lm</math> ("the cost")</p>
phrases 2 & 3	每吨 海水淡化 处理 的 成本 在 5 元 的 基础 上 进 一 步 下 降 。
	<p>the cost <math>p_1</math></p> <p>of <math>p_2</math></p> <p>seawater desalination treatment <math>p_3</math></p> <p><b>Skeleton:</b> <math>g(d_\tau; \mathbf{w}^\tau, m, lm^\tau) = \mathbf{w}_m^\tau \cdot \mathbf{f}_m(p_1) + w_{lm}^\tau \cdot lm^\tau</math> ("the cost X")</p> <p><b>Full:</b> <math>g(d; \mathbf{w}, m, lm) = \mathbf{w}_m \cdot \mathbf{f}_m(p_1 \circ p_2 \circ p_3) + w_{lm} \cdot lm</math> ("the cost of seawater desalination treatment")</p>
phrases 4 & 5	每吨 海水淡化 处理 的 成本 在 5 元 的 基础 上 进 一 步 下 降 。
	<p>the cost <math>p_1</math></p> <p>of <math>p_2</math></p> <p>seawater desalination treatment <math>p_3</math></p> <p>has been further <math>p_4</math></p> <p>reduced <math>p_5</math></p> <p><b>Skeleton:</b> <math>g(d_\tau; \mathbf{w}^\tau, m, lm^\tau) = \mathbf{w}_m^\tau \cdot \mathbf{f}_m(p_1 \circ p_4 \circ p_5) + w_{lm}^\tau \cdot lm^\tau</math> ("the cost X has been further reduced")</p> <p><b>Full:</b> <math>g(d; \mathbf{w}, m, lm) = \mathbf{w}_m \cdot \mathbf{f}_m(p_1 \circ p_2 \circ \dots \circ p_5) + w_{lm} \cdot lm</math> ("the cost of seawater ... further reduced")</p>
phrases 6-9	每吨 海水淡化 处理 的 成本 在 5 元 的 基础 上 进 一 步 下 降 。
	<p>the cost <math>p_1</math></p> <p>of <math>p_2</math></p> <p>seawater desalination treatment <math>p_3</math></p> <p>has been further <math>p_4</math></p> <p>reduced <math>p_5</math></p> <p>from <math>p_6</math></p> <p>5 yuan <math>p_7</math></p> <p>per ton <math>p_8</math></p> <p><math>p_9</math></p> <p><b>Skeleton:</b> <math>g(d_\tau; \mathbf{w}^\tau, m, lm^\tau) = \mathbf{w}_m^\tau \cdot \mathbf{f}_m(p_1 \circ p_4 \circ p_5 \circ p_9) + w_{lm}^\tau \cdot lm^\tau</math> ("the cost X has been further reduced X.")</p> <p><b>Full:</b> <math>g(d; \mathbf{w}, m, lm) = \mathbf{w}_m \cdot \mathbf{f}_m(p_1 \circ p_2 \circ \dots \circ p_9) + w_{lm} \cdot lm</math> ("the cost of seawater ... per ton.")</p>

# 実験

- 原語対: 中英
- SMTシステム: NiuTrans
- データ:
  - パラレル・コーパス: 2.7M 文対
  - 言語モデル用: (上記に付け加え) Gigaword corpus の Xinhua 箇所
- スケルトン言語モデル
  - パラレルコーパスの目的原語側の非重要箇所をXとし学習
  - スケルトン・システム: t3 toolkit
- 評価セット:
  - MT04, MT05で利用された評価セットを利用

# 評価

system	Entry		MT06 (Dev)		MT04		MT05		All	
	dev-skel	test-skel	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
baseline	-	-	35.06	60.54	38.53	61.15	34.32	62.82	36.64	61.54
SBMT	manual	manual	<b>35.71</b>	<b>59.60</b>	38.99	<b>60.67</b>	<b>35.35</b>	<b>61.60</b>	<b>37.30</b>	<b>60.73</b>
SBMT	manual	auto	<b>35.72</b>	<b>59.62</b>	38.75	61.16	35.02	<b>62.20</b>	37.03	61.19
SBMT	auto	auto	35.57	<b>59.66</b>	<b>39.21</b>	<b>60.59</b>	<b>35.29</b>	<b>61.89</b>	<b>37.33</b>	<b>60.80</b>
$-lm^T$	auto	auto	35.23	<b>60.17</b>	38.86	60.78	34.82	<b>62.46</b>	36.99	61.16
$-m^T$	auto	auto	35.50	<b>59.69</b>	39.00	<b>60.69</b>	<b>35.10</b>	<b>62.03</b>	37.12	<b>60.90</b>
s-space	-	-	35.00	60.50	38.39	61.20	34.33	62.90	36.57	61.58
s-feat.	-	-	35.16	60.50	38.60	61.17	34.25	62.88	36.70	61.58