A Novel Dependency-to-String Model for Statistical Machine Translation

Jun Xie    Haitao Mi    Qun Liu
Institute of Computing Technology
Chinese Academy of Sciences
Dependency Structure

举行 (hold)

世界杯 (World Cup)

在 (in/at) 成功 (successfully)

2010年 (2010) FIFA (FIFA)

南非 (South Africa)
Dependency Structure

- Best phrasal cohesion properties (Fox, 2002)
Dependency Structure

- Best phrasal cohesion properties (Fox, 2002)
- Dependencies encode the semantic relations directly
Dependency Structures in SMT
Dependency Structures in SMT

string-to-dependency

(Shen et al., 2008)
Dependency Structures in SMT

string-to-dependency

dependency-to-string

Bush held a talk with Sharon

bushi yu shalong juxing le huitan

(Shen et al., 2008)

(Xiong et al., 2007)
Dependency Structures in SMT

- **string-to-dependency**
  - (Shen et al., 2008)

- **dependency-to-string**
  - (Xiong et al., 2007)

- **dependency-to-dependency**
  - (Quirk et al., 2005; Ding et al., 2005; Lin et al., 2004)
Dependency Structures in SMT

(Xiong et al., 2007)
Dependency Structures in SMT

*(Xiong et al., 2007)*

- Treelets as elementary structures
Dependency Structures in SMT

- Treelets as elementary structures
- Heuristics to assist translations

(Xiong et al., 2007)
Dependency Structures in SMT

- Treelets as elementary structures
- Heuristics to assist translations
- Problem: heuristics is not reliable

(Xiong et al., 2007)
Dependency Structures in SMT

- Treelets as elementary structures
- Heuristics to assist translations
- Problem: heuristics is not reliable
- Idea: specify ordering directly in translation rules

(Xiong et al., 2007)
Preview of Results

- We specify ordering information directly in head-dependents rules
Preview of Results

- We specify ordering information directly in head-dependents rules
Preview of Results

- We specify ordering information directly in head-dependents rules

- No heuristics
Preview of Results

- We specify ordering information directly in head-dependents rules

- No heuristics
- Better performance
The first source dependency-based model outperforming the state-of-the-art models
Our dependency-to-string model captures patterns of sentences and phrases.
Intuitions
Intuitions

- Intuition 1
  - A sentence is formed by
    - An instance of sentence pattern
    - Instances of phrase patterns
Sentence pattern

![Diagram with NR, P, and AD connected to VV]
世界杯在成功举行

Sentence pattern
Sentence pattern

Phrase patterns
2010年FIFA世界杯在成功举行

Sentence pattern

Phrase patterns
Sentence pattern

Phrase patterns

Sentence: 2010年FIFA世界杯在成功举行

Phrase patterns:

Phrase: 2010年

Phrase: FIFA

Phrase: 世界杯
Sentence pattern

Phrase patterns

Sentence: 在南非世界杯成功举行。

Phrases: 南非, 世界杯, 在, 成功, 举行.
2010年 FIFA 世界杯 在 南非 成功 举行
2010niànn FIFAShìjièběizàiNánfēichénggōngjǔxíng
Elementary structures: Head-dependents relations

2010年 FIFA 世界杯 在 南非 成功 举行
2010niàn shìjièbēi zài nánfēi chénggōng jǔxíng
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
  - A pattern has many variants
Intuitions

- Intuition 2
- A pattern has many variants
Grammar

- Head-dependents rules

世界杯

在

成功

举行

jǔxíng

X1 was held successfully X2

shìjìébēi

zài

chénggōng
Grammar

- Head-dependents rules

世界杯 in X1 was held X3 in X2

Grammar diagram:

世界杯 举行 世界杯

在 举行 举行

AD3 举行 举行

shijiebei zai
Grammar

- Head-dependents rules
Grammar

- Head rules

成功 ➔ successfully
chénggōng
Grammar

• Derivation

2010年 FIFA 世界杯 在 南非 成功 举行
2010年/NTFIFA/NR

世界杯/NR

在/P

南非/NR

成功/AD

举行/VV
世界杯2010年在南非成功举行。
世界杯在2010年在南非举行，成功举行。

R1: 世界杯 was held in South Africa.
世界杯在南非举行。2010年南非世界杯成功举行。
2010年/FIFA/世界杯成功在南非举行。
2010 FIFA World Cup was held successfully in South Africa in 2010.
2010 FIFA World Cup was held successfully in South Africa in 2010.
The 2010 FIFA World Cup was held successfully in South Africa.
2010 FIFA World Cup was held in South Africa.
2010 FIFA World Cup was held in South Africa successfully.
2010 FIFA World Cup was held successfully in South Africa.

R3:
成功 → successfully
2010 FIFA World Cup was held successfully in South Africa.

R3: 成功 → successfully
2010 FIFA World Cup was held successfully in 南非/NR.
The 2010 FIFA World Cup was held successfully in Namibia.
2010 FIFA World Cup was held successfully in 南非/NR.
2010 FIFA World Cup was held successfully in South Africa.
2010 FIFA World Cup was held successfully in 南非/NR.
2010 FIFA World Cup was held successfully in 南非/NR.

R5:
南非 → South Africa
2010 FIFA World Cup was held successfully in 南非/NR

R5:
南非  ➔ South Africa
2010 FIFA World Cup was held successfully in South Africa.
2010 FIFA World Cup was held successfully in South Africa.
Decoding

- Bottom-up chart parsing
Decoding

- Bottom-up chart parsing
- Post-order transverse
世界杯

举行/VV

2010年/NTFIFA/NR

在/P

南非/NR

成功/AD
2010年/FIFA在南非成功举行世界杯。
2010年

2010年

举行

世界杯

南非

在

成功

FIFA
2010年

Rule set

2010年 → 2010

...
2010年/FIFA/世界杯在南非举行成功。
2010年/FIFA/南非举行世界杯/成功

世界杯/在南非举行

世界杯/2010年/FIFA/NR

世界杯/NR/NT/NR
2010年FIFA世界杯在南非举行成功
2010年FIFA世界杯
举行
在
成功
南非

Rule set

世界杯
2010年
FIFA
World Cup

X1 X2 X3

NT1 NR2

NT
NR
Experiments

Large scale experiments on Chinese-to-English translation
**Baselines**
- Constituency-to-String model (Liu et al., 2006)
- Replication of HPB (Chiang, 2005)

**Data**
- Training: 1.5M sentence pairs from LDC
- LM: Xinhua portion of Gigaword
- Dev Set: NIST 2002
- Test Sets: NIST 2004 /2005
- Stanford Parser (Klein et al., 2003)
Rule #

![Bar chart showing the distribution of Rule # (M) with categories cons2str, hiero-re, and dep2str. The chart indicates that cons2str has 30 occurrences, hiero-re has 148 occurrences, and dep2str has 56 occurrences.]
The first source dependency-based model outperforming the state-of-the-art models
Analysis of Translations
Analysis of Translations

- Better long distance reordering
After a brief talk with Powell at the US State Department, Barnier said...
After a brief talk with Powell at the US State Department, Barnier said

Reference:
After a brief talk with Powell at the US State Department, Barnier said

Cons2str:
Barnier after brief talks in US State Department and Powell said

Hiero-re:
After a short meeting with Barnier on the US State Department, Powell said
Example 1

After a brief talk with Powell at the US State Department, Barnier said

Reference:
After a brief talk with Powell at the US State Department, Barnier said

Cons2str:
Barnier after brief talks in US State Department and Powell said

Hiero-re:
After a short meeting with Barnier on the US State Department, Powell said

Dep2str
After brief talks with Powell, the US State Department Barnier said
Example 2

China appreciates the efforts of the Counter-Terrorism Committee to promote…
China appreciates the efforts of the Counter-Terrorism Committee to promote…

Cons2str:
China appreciates the efforts, building for the efforts,

Hiero-re:
China appreciates support for efforts
China appreciates the efforts of the Counter-Terrorism Committee to promote…

**Dep2str:**
China appreciates efforts of Anti-Terrorism …
Conclusion

- The first source dependency-based model that outperforms state-of-the-art models
Conclusion

- The first source dependency-based model that outperforms state-of-the-art models
- Better long distance reordering
Conclusion

- The first source dependency-based model that outperforms state-of-the-art models
- Better long distance reordering
- No heuristics
Conclusion

● The first source dependency-based model that outperforms state-of-the-art models

● Better long distance reordering

● No heuristics

● Future work
  ● Employ parse forest to alleviate parse errors
Conclusion

- The first source dependency-based model that outperforms state-of-the-art models
- Better long distance reordering
- No heuristics
- Future work
  - Employ parse forest to alleviate parse errors
  - Exploit typed dependencies to direct translation
Thanks!
About heuristics

The 2010 FIFA World Cup was held successfully in South Africa.
About heuristics

2010 … Cup was held successfully in South Africa
Rule acquisition

- Intuition 3
  - Semantically close items tend to stay close across languages
2010年 FIFA 世界杯 在 南非 成功 举行

2010 FIFA World Cup was held successfully in South Africa
2010 FIFA World Cup was held successfully in South Africa.
2010 FIFA [World Cup] was held successfully in [South Africa]
Rule acquisition

- Acceptable head-dependents fragments
  - Intuition 3
- Rule induction
  - Intuition 2
2010 FIFA [World Cup] was held successfully in [South Africa].
2010年FIFA世界杯在南非成功举行。
世界杯在南非成功举行。
Rule acquisition

- Rule induction
  - Generate variants of the encoded pattern
Rule acquisition

- Rule induction
  - Generate variants of the encoded pattern
Rule acquisition

- Rule with four lexical words (fully lexicalized)

世界杯

在

成功

举行

→

X1 held successfully X2
Rule acquisition

- Rules with three lexical words

---

世界杯

在

成功

VV4

→

X1  X4  successfully  X2  ...

---
Rule acquisition

- Rules with two lexical words

\[ \text{NR}_1 \rightarrow \text{成功} \rightarrow \text{举行} \rightarrow \text{X1 held successfully X2} \ldots \]
Rule acquisition

- Rules with one lexical word

NR_1 ↓ P_2 ↓ AD_3

举行 → X1 held X3 X2 ...

...
Rule acquisition

- Rule of the pattern
  (fully unlexicalized)
About restrictions

- Three types of nodes
  - Head node
  - Internal nodes
  - Leave nodes
About restrictions

- Open word class
  - \{CD,DT,OD,JJ,NN,NR,NT,AD,FW,PN\}
Why not multiple layer?

- Head dependents rules
  - Better long distance relation capture capability
Why not multiple layer?

- Head dependents rules
- Better long distance relation capture capability
2010年FIFA世界杯在南非成功举行。
2010年 FIFA 世界杯 在 南非 成功 举行
Why not multiple layer?

- Data sparseness problem
Why not multiple layer?

- Data sparseness problem
  - Head-dependents relation with verb as head
  - More than 4 nodes
Why not multiple layer?

- Data sparseness problem
  - Head-dependents relation with verb as head
    - More than 4 nodes
- Massive grammar size
Why not GHKM span?

- Top level head-dependents relations
  - the GHKM spans often overlap
2010年 FIFA [World Cup] was held successfully in [South Africa].
2010年FIFA在南非世界杯成功举行。
Why not GHKM span?

- A looser constraint under the intuition
  --semantically closed items tends to stay close across languages