Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents

Dun Deng, Nianwen Xue, Shiman Guo

Brandeis University

2015-06-04

Denver, Colorado
What’s an ideal language resource for statistical machine translation?

- Lots and lots of parallel texts?
- A large word and phrase aligned parallel treebank?
- A large semantically annotated parallel corpus?
  - Parallel Abstract Meaning Representation Corpus
- Other
What’s an ideal language resource for statistical machine translation?

• Lots and lots of parallel texts?
✓ A large word and phrase aligned parallel treebank?

• A large semantically annotated parallel corpus?
  – Parallel Abstract Meaning Representation Corpus

• Other
Issues with treebanks and word alignments

• Status quo:
  – Manually annotated treebanks and word alignments independently conceived
    • Inevitable incompatibilities between word alignment and the syntactic structure
  – Treebanks on both sides are independently conceived
    • Incompatibilities between the parse trees of a sentence pair
  – Treebanks not optimized for MT
    • Trees too shallow, too deep

• Frustrated MT users
  – Forget about parse trees!
  – Or use automatically induced parse trees!
Addressing this problem

- Create a hierarchically aligned Chinese English parallel treebank that
  - harmonizes word alignment and the syntactic structure of a sentence,
  - synchronizes the parse trees of the sentence pair
  - Empirically determine the amount of structure that is needed during alignment
Word alignment is deceptively complicated

✓ An equivalent exists in the target-language sentence, which matches the word in both lexical meaning and grammatical function

? There is a candidate in the target-language sentence, which does not have the same lexical meaning and/or grammatical function as the word but could be aligned in the given context

? The word has no translation counterpart in the target-language sentence at all.
Word alignment

laoshi  ba  xuesheng  biaoyang  le
老师  把  学生  表扬  了
the  teacher  praised  the  student

the teacher praised the student.
Word alignment

laoshi  
老师

ba  
把

xuesheng  
学生

biaoyang  
表扬

le  
了

the  
the

teacher  
teacher

praised  
praised

the  
the

student  
student

老师 ⇔ teacher  
老师 ⇔ the teacher

学生 ⇔ student  
学生 ⇔ the student

把学生 ⇔ student  
把学生 ⇔ the student
“Syntactic annotation” in word alignment

The teacher praised the student.

laoshi  ba  xuesheng  biaoyang  le
老师  把  学生  表扬  了
the  teacher  praised  the  student
the

Brandeis University
Redundancy and conflict with parse trees
Causes of conflict

• Word alignment standards and treebanking standards are independently established and are meant to be used separately
• Need an approach that systematically considers the interaction between word alignment and syntactic structure to maximize the utility of aligned parse trees
• Solution: hierarchical alignment
  – “sure” alignments only at word level, others aligned indirectly via node alignment in their context
Hierarchical alignment

laoshi  
老师  
the teacher

ba  
把  
praised  
the student

xuesheng  
学生  
NP

biaoyang  
表扬  
le  
了

12
老师把学生表扬了

laoshi  

the teacher

ba  

the 

xuesheng  

praised

biaoyang  

student

le  

了
Hierarchical alignment

laoshi 老师
ba 把
xiaosheng 学生
biaoyang 表扬
le 了

the teacher praised the student.
The teacher praised the student.
Issues in aligning parse trees

• Too much structure
  – Not all nodes in a parse tree are aligned or alignable.
  – This might suggest that not all the syntactic structures annotated in existing treebanks are necessary for MT purposes

• Too little structure
  – Flat structures in the parse trees can prevent legitimate alignments
  – Needs to revise trees in order to get proper alignments
The Senate reopened the budget talks last week.
Potential benefits of hierarchical alignment

• Captures unconstrained long-distance lexical dependencies
• Allows the extraction of linguistically interpretable rules
• Smaller but equally potent translation table
• Better MT accuracy?
Fewer rules
Long distance dependency: Passives

[可 禁止 X₁ X₂ <---> X₁ is prohibited from X₂]
Long distance dependency: 把 construction

[把 X 定为非法因素 <-> outlaw X]
Long distance dependency: discourse Connectives

[将 X_1，这样 X_2 \iff \text{will } X_1 \text{ so that } X_2]
Long distance dependency: Questions

[ $X_1$ 为什么 不 $X_2$ --- Why do n’t $X_1 X_2$]
Alignment in syntactic context: Conjunction

\[[X, Y <> X \text{ and } Y]\]
Conclusions and future work

• Described a hierarchical alignment approach that systematically considers the interaction between word alignment and the syntactic structure of a sentence.

• Showed that such alignments can be used to extract translation rules that cover long-distance dependencies.

• Aligned 10K sentence pairs annotated with PTB and CTB trees

• To do: revising treebanking guidelines and synchronizing the parse trees

• More info can be found at:
  – http://www.cs.brandeis.edu/~clp/ctb/hacept/
Acknowledgements

• Libin Shen for inspirations
• Tool support by the IBM Multilingual NLP Technologies Group: Niyu Ge, Abe Ittycheriah, Salim Roukos
• Annotators: Gao Hui, Guo Shiman, Wang Tse-Ming, Zhou Linya
• Funded by DARPA BOLT Program via IBM