Clause Restructuring for Statistical Machine Translation

Michael Collins (MIT CSAIL)
Philipp Koehn (University of Edinburgh)
Ivona Kučerová (MIT Linguistics)

Outline

- Current statistical machine translation (SMT) systems make limited use of syntax
- Problems when two languages have differing word order (e.g., German $\Rightarrow$ English translation)
- Our solution: pre-processing step that

  re-orders source language sentences
- Result: **A simple, direct method for syntax within a phrase-based SMT system**

A Problem in Translation from German to English

Original sentence:

Ich werde Ihnen den Report aushaendigen...

English gloss:

I will to_you the report pass_on...

A Problem in Translation from German to English

Original sentence:

damit Sie den eventuell uebernehmen koennen.

English gloss:

so_that you that perhaps adopt can.

A Sketch of Our Approach

Step 1:

Reorder the source language (German)

Ich werde Ihnen den Report aushaendigen,
damit Sie den eventuell uebernehmen koennen.

⇒ Ich werde aushaendigen Ihnen den Report,
damit Sie koennen uebernehmen den eventuell

(I will pass_on to_you the report, so_that you can adopt that perhaps)

Step 2:

Apply a phrase-based system to the re-ordered input

Motivation

- Phrase-based models offer a relatively simple way to handle differing word order
- Linguistically, we have a good idea of the differences between different languages
- A set of 6 simple, linguistically motivated rules gives significant improvements

Overview

- Linguistic motivation
  - German sentence structure
  - Free word order
- Phrase-based systems
- A re-ordering strategy
- Experiments
- Related work
- Future work

German Sentence Structure

Matrix clauses:

- Finite verb in 2nd position
- Non-finite verbal forms clause final

Example:

Ich werde Ihnen den Report aushaendigen

I will to_you the report pass_on

German Sentence Structure

Subordinate clauses:

- Finite verb is in clause final position
- Non-finite verb precedes the finite verb

Example:

damit Sie den eventuell uebernehmen koennen

so_that you that perhaps adopt can

Free Word Order

Example: *any* constituent (not only subject) can appear in first position

I will to_you the report pass_on

To_you will I the report pass_on

The report will I to_you pass_on

Overview

- Linguistic motivation
- Phrase-based systems
- A re-ordering strategy
- Experiments
- Related work
- Future work

Phrase-based Systems

(E.g., Och et. al, 1999; Marcu & Wong, 2002; Koehn et al., 2003)

Phrasal lexicon induced from IBM alignment, e.g.:

I will	↔	Ich werde
to you	↔	Ihnen
the report	↔	den Report
pass on	↔	aushaendigen

Greedy Decoding

- Greedy, left-to-right decoding method:
  1. At each step, translate a German “phrase”
  2. Each step has an associated cost (probability)

~~Ich~~ ~~werde~~ Ihnen den Report aushaendigen

I will pass on

Difficulties for Phrase-based Systems

- In practice, high cost for skipping words
- Some natural collocations are not contiguous, e.g.:

wir machen die Tuer auf

(we open the door)

'machen ... auf' $\Rightarrow$ *to_open*

BUT: 'machen' $\Rightarrow$ *to_make*

Overview

- Linguistic motivation
- Phrase-based systems
- A re-ordering strategy
- Experiments
- Related work
- Future work

A Sketch of Our Approach

Step 1: Reorder the source language

Ich werde Ihnen den Report aushaendigen...
damit Sie den eventuell uebernehmen koennen.

⇒ Ich werde aushaendigen Ihnen den Report,
damit Sie koennen uebernehmen den eventuell

(I will pass_on to_you the report, so_that you can adopt that perhaps)

Step 2: Apply a phrase-based system to the re-ordered input

Reordering Process

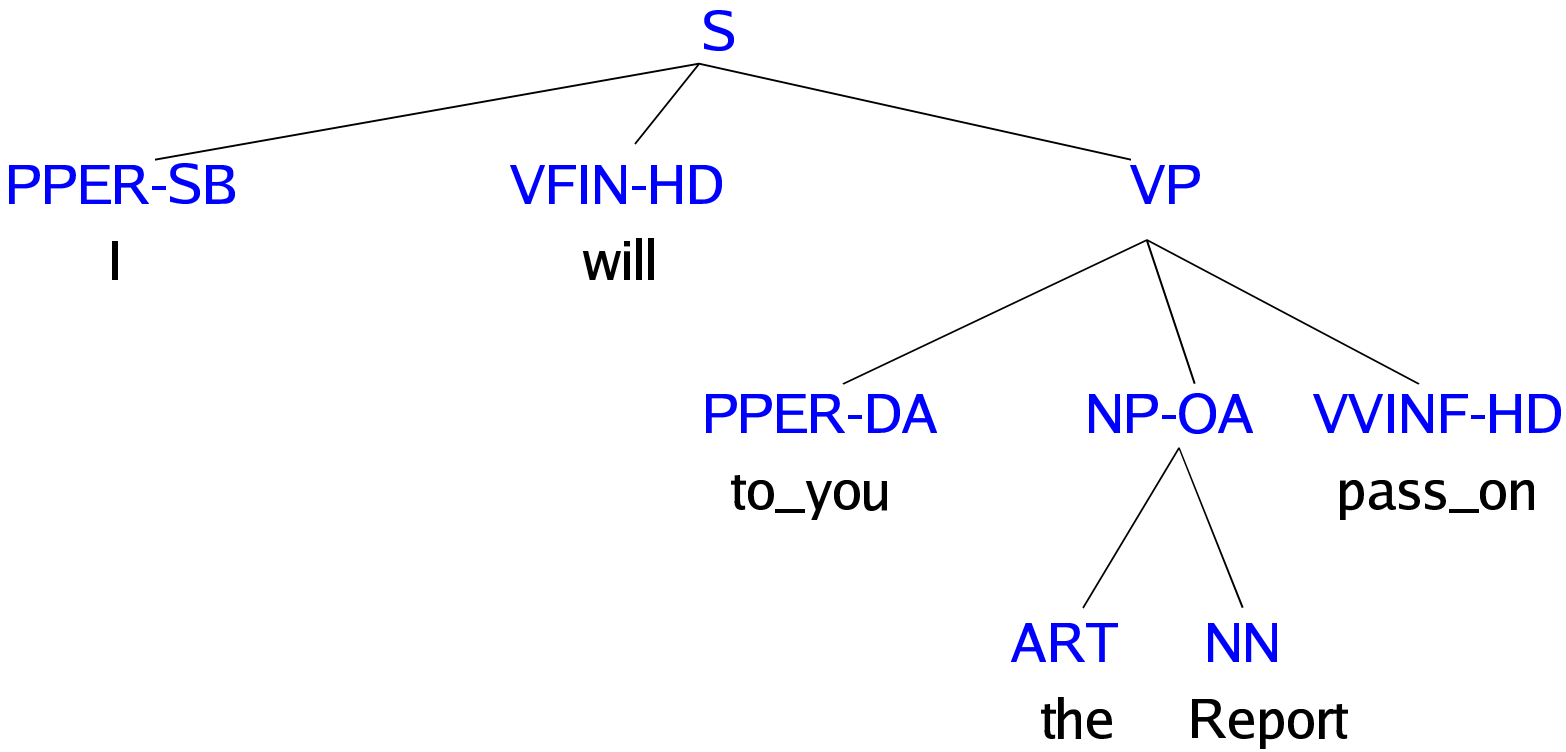
- Step 1:

  Parsing source language sentences by the parser described in Dubey and Keller, 2003

- Step 2:

  Applying 6 syntactic transformations (moving verbs, subjects, particles and negation)

Example of A Parse Tree



Clause Restructuring

1. Verb initial:

Within a VP, move the head to the initial position

VP-OC	PDS-OA den	*that*
	ADJD-MO eventuell	*perhaps*
	VVINF-HD uebernehmen	*adopt*
VMFIN-HD	koennen	*can*

⇒

VP-OC	VVINF-HD uebernehmen	*adopt*
	PDS-OA den	*that*
	ADJD-MO eventuell	*perhaps*
VMFIN-HD	koennen	*can*

Clause Restructuring

2. Verb 2nd:

In a subordinated clause, move the head of the clause to follow the complementizer

S-MO	KOUS-CP	damit	*so that*
	PPER-SB	Sie	*you*
	VP-OC	VVINF-HD uebernehmen	*adopt*
		...	
	VMFIN-HD	koennen	*can*

⇒

S-MO	KOUS-CP	damit	*so that*
	VMFIN-HD	koennen	*can*
	PPER-SB	Sie	*you*
	VP-OC	VVINF-HD uebernehmen	*adopt*
		...	

Clause Restructuring

3. Move subject:

The subject is moved to directly precede the head of the clause

S-MO	KOUS-CP damit	*so that*
	VMFIN-HD koennen	*can*
	PPER-SB Sie	*you*
	VP-OC VVINF-HD uebernehmen	*adopt*
	...	

⇒

S-MO	KOUS-CP damit	*so that*
	PPER-SB Sie	*you*
	VMFIN-HD koennen	*can*
	VP-OC VVINF-HD uebernehmen	*adopt*

Clause Restructuring

4. Particles:

In verb particle constructions, the particle is moved to precede the finite verb

S	PPER-SB	Wir	*we*
	VVFIN-HD	fordern	**accept**
	NP-OA ART	das	*the*
	NN	Praesidium	*presidency*
	PTKVZ-SVP	auf	**PARTICLE**

⇒

S	PPER-SB	Wir	*we*
	PTKVZ-SVP	auf	--
	VVFIN-HD	fordern	*accept*
	NP-OA ART	das	*the*
	NN	Praesidium	*presidency*

Clause Restructuring

5. Infinitives:

Infinitives are moved to directly follow the finite verb within a clause

S	PPER-SB Wir	*we*
	VMFIN-HD konnten	*could*
	PPER-OA es	*it*
	PTK-NEG nicht	*not*
	VP-OC VVINF-HD einreichen	*submit*
	...	

⇒

S	PPER-SB Wir	*we*
	VMFIN-HD konnten	*could*
	VVINF-HD einreichen	*submit*
	PPER-OA es	*it*
	PTK-NEG nicht	*not*
	VP-OC ...	

Clause Restructuring

6. Negation:

Negative particle is moved to directly follow the finite verb

S	PPER-SB Wir	*we*
	VMFIN-HD konnten	*could*
	VVINF-HD einreichen	*submit*
	PPER-OA es	*it*
	PTK-NEG nicht	*not*
⇒		
S	PPER-SB Wir	*we*
	VMFIN-HD konnten	*could*
	PTK-NEG nicht	*not*
	VVINF-HD einreichen	*submit*
	PPER-OA es	*it*

Overview

- Linguistic motivation
- Phrase-based systems
- A re-ordering strategy
- Experiments
- Related work
- Future work

Experiments

- Data: Europarl corpus
- 751,088 sentence pairs (15,256,792 German words, 16,052,269 English words)
- Translation performance measured on a 2000 sentence test set from Europarl
- Average sentence length of 28 words

Experiments

1. Underlying system:
   - Phrase-based method of Koehn et al. 2003

3. Baseline method:
   - No reordering on training or test data

4. Reordering method:
   - Reordered training data, trained system
   - Reordered test data before applying system

Results

- BLEU score (Papineni et al., 2002):
  - baseline system: 25.2% BLEU
  - the new system: 26.8% BLEU

- Significant at $p<0.01$ using the sign test
  (*see paper for details*)

Human Translation Judgments

- 100 sentences (10 to 20 words in length)
- Two annotators
- Three versions:
  - Human translation, baseline translation, reordered translation
- Judgments: Worse/better or equal
- Sentences were chosen at random, baseline and reordered translations presented in random order

Human Translation Judgments

	+	=	–
Annotator 1	40 %	40 %	20 %
Annotator 2	44 %	37 %	19 %

- + = reordered translation better
- – = baseline better
- = = equal

Example: Reordered System Better

H: i think it is wrong in principle to have such measures in the european union.

R: i believe that it is wrong in principle to take such measures in the european union.

B: i believe that it is wrong in principle, such measures in the european union to take.

H = human translation, **R** = reordered, **B** = baseline

(*see paper for randomly chosen examples*)

Example: Reordered System Better

H: the current difficulties should encourage us to redouble our efforts to promote cooperation in the euro-mediterranean framework

R: the current problems should spur us to intensify our efforts to promote cooperation within the framework of the e-prozesses

B: the current problems should spur us, our efforts to promote cooperation within the framework of the e-prozesses to be intensified

Example: Reordered System Better

H: to go on subsidising tobacco cultivation at the same time is a downright contradiction in terms

R: at the same time, to continue to subsidise tobacco growing is schizophrenic.

B: at the same time, continue to subsidise tobacco growing, it is quite schizophrenic.

Example: Reordered System Better

H: we have voted against the report by mrs lalumiere for reasons that include the following

R: we have voted, amongst other things, for the following reasons against the report by mrs lalumiere

B: we have, among other things, for the following reasons against the report by mrs lalumiere voted

Overview

- Linguistic motivation
- Phrase-based systems
- A re-ordering strategy
- Experiments
- Related work
- Future work

Research on Syntax-based SMT

- Syntax-to-syntax SMT (Alshawi, 1996; Wu, 1997; Gildea, 2003; Melamed, 2004)
- Syntax on target language alone (Yamada and Knight, 2001)
- Pre-processing input to phrase-based systems (Berger et al., 1996; Niessen and Ney, 2004; Xia and McCord, 2004)

Our Approach vs. Xia and McCord, 2004

- Xia and McCord, 2004 automatically learn rules for French ⇒ English
- Rules are stated as context-free transformations
- Contrasts:
  - German is a language with free word order, and has greater word order differences from English than French
  - We require 6 transformations, vs. 56,000 learned rules for Xia and McCord
- Future work should compare and combine the two approaches

Future Work

- Other languages
- Automatic learning of rules of this type
- Combining the method with other approaches
  - E.g. Xia and McCord's method, Chiang's method (ACL 05)
  - Do the two methods complement each other?
- Split learning of the rules into two parts:
  - Learning on a source language only (monolingual, bigger corpora)
  - Learning on a pair of languages