Evaluating Translational Correspondence using Annotation Projection

R. Hwa, P. Resnik, A. Weinberg & O. Kolak (2002)

Presented by Jeremy G. Kahn

Presentation for Ling 580 (Machine Translation)
10 Jan 2006

Introduction

The Main Issue: syntactic divergence

Trees in 2 languages may be homomorphic. . . or not

- same basic shape, different rotations of CFG
- different basic shape of CFG (rules can't correspond)

Direct Correspondence Assumption (DCA):

"the syntactic relationships in one language directly map to the syntactic relationships in the other."

In this paper, "syntactic relationships" $\rightarrow$ dependencies.

Exploring the DCA

DCA is implicit in:

1. Inversion Transduction Grammar (ITG) (& D. Wu's SITG)

   Suppose:

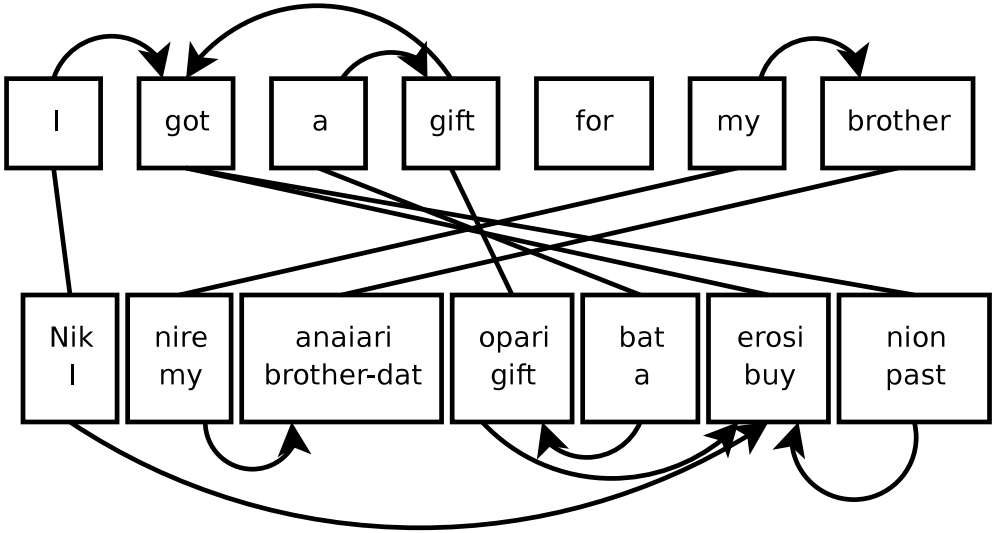
   L1 is SVO, D-N language; L2 is SOV, N-D language

   - S $\rightarrow$ NP VP
   - NP $\rightarrow$ [D N]
   - VP $\rightarrow$ [V NP]

   Note special meaning for bracketing.

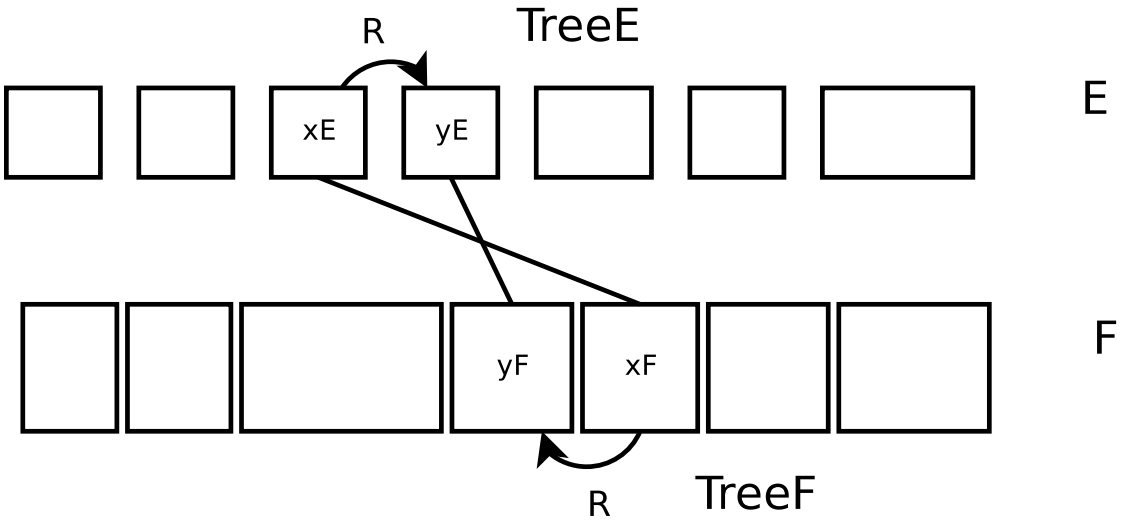
   not mentioned in the paper: Melamed's synchronous CFGs, a superset of ITG

2. synchronized dependency trees (Alshawi et al. 2000)


I
got
a
gift
for
my
brother
Nik
I
nire
my
anaiari
brother-dat
opari
gift
bat
a
erosi
buy
nion
past

DCA as formalism

Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures Tree$_E$ and Tree$_F$ if nodes x_E and y_E of Tree$_E$ are aligned with nodes x_F and y_F of Tree$_F$ respectively, and if syntactic relationship $R(x_E, y_E)$ holds in Tree$_E$ then $R(x_F, y_F)$ holds in Tree$_F$



Why is the DCA good?

matches a linguistic thought: thematics (dependencies) are held constant but word order may change

fairly elegant conceptually

allows us to take advantage of formalisms like ITG, synchronized trees

Potential problems with the DCA

1. word-to-word correspondence questions - morphology in one language may be word (or word-order) in another
   - the Basque dative vs. English 'for'
   - Basque 'buy', 'past' two words vs. English 'bought' portmanteau
2. tree structures in use may not be the right rotational operations (not mentioned in the text): SVO vs OSV languages (Arabic), using 2-branching

   ex: [I [like apples]] vs [apples [I like]]

   VP relation becomes disconnected

Looking at the DCA: a task

Comparing English (En) and Chinese (Zh) structures through projection

Given: Gold English parses (dependencies) & gold word-alignment

Task: project En (dependency) structures onto Zh word sequence

Evaluate: projected En$\rightarrow$Zh dependencies vs. independently derived Zh dependencies (unlabeled dependency P,R,F)

Corpus

Dev set:

124 Zh sentences (av length 23.7), En translations by hand.

Zh dep trees derived by hand (guided by TB). (2 annotators, 92.4% annotation agreement)

Test set:

88 Zh sentences (av length 19.0), En translations from NIST MT project

Zh dep trees derived automatically from TB (a la Xia & Palmer 2001)

Both sets:

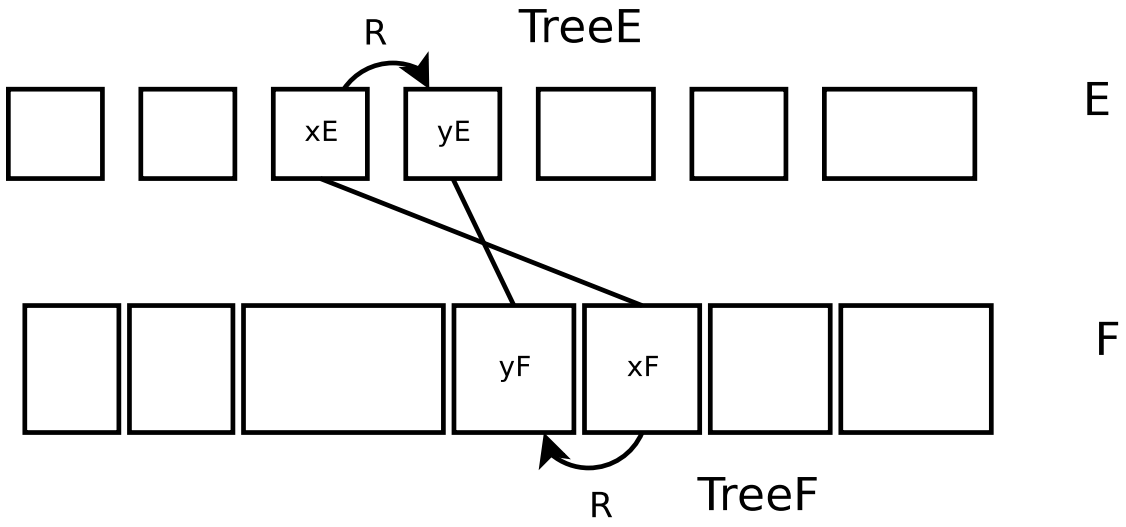
Zh trees originate with Zh treebank (but deps derived differently)

En deps generated via parse (Collins 97) and hand-correct

Algorithm 1: Direct Projection Algorithm (DPA)

4 cases:

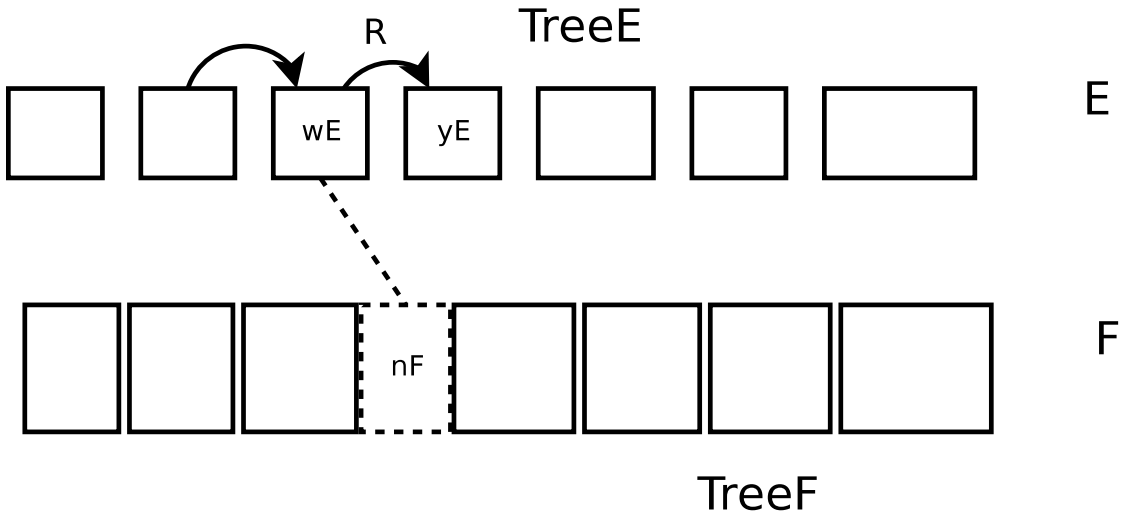
- paired 1-to-1 alignments: two 1-1 alignments that share an E-side dependency $\rightarrow$ induce an F-side dependency.



Algorithm 1: Direct Projection Algorithm (DPA) #2

Unaligned E-side:

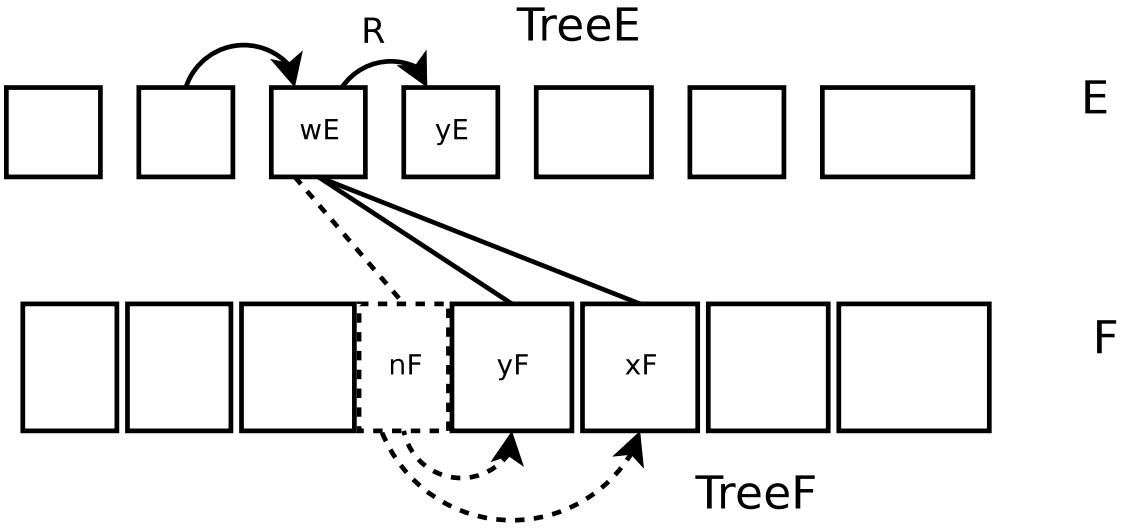
- En words w_e with no Zh word: create an F-side word n_f. For each E-side dependency involving w_e, if the non-w_e token (x_e) aligns 1-to-1 with an F-side word (x_f) induce an F-side dependency between n_f and x_f.



Algorithm 1: Direct Projection Algorithm (DPA) #3

1 En to many Zh:

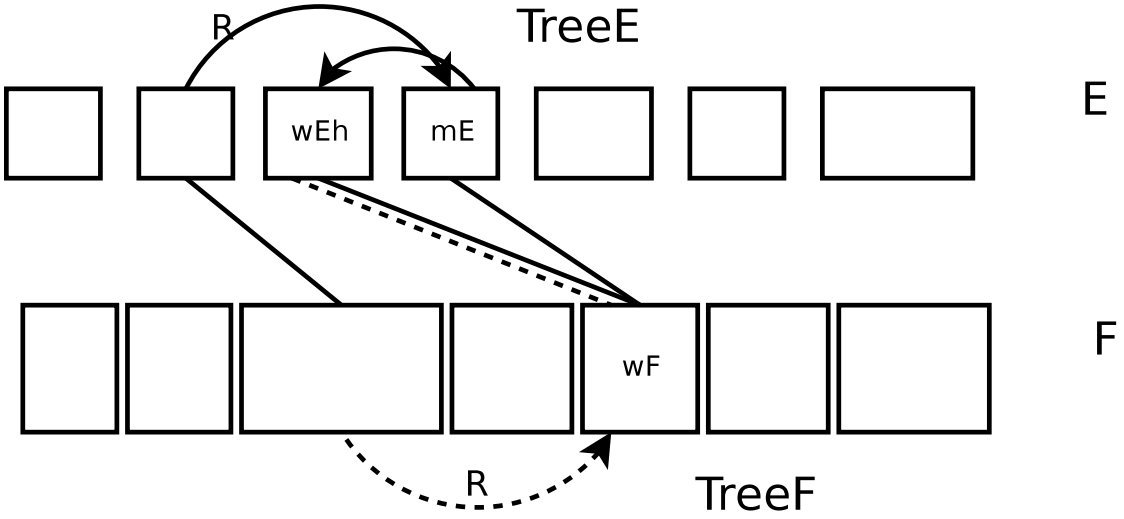
- A single E-side word w_E aligned with several w_f words: invent an F-side word n_f and make all the w_f children of that word. Align w_e to n_f. (presumably, return to case 1)



Algorithm 1: Direct Projection Algorithm (DPA) #4

many En to 1 Zh:

- A single F-side word w_f is aligned with several w_e words. (Select a head w_{eh} from w_e words), align w_f with w_{eh} only. Also, any dependencies that involve the modifier (non-head) E-side words (m_e) should be pointed at w_f on the F-side.



Many-to-many (vaguely) is 1-to-many then many-to-1 (?)

Error analysis of DPA

Dev set results ("exp 1") show that DPA on dev set is lousy: P 30.1, R 39.1

Error analysis: lots of multiply-aligned, unaligned tokens.

In particular, difference in morph boundaries and word content.

- Chinese measure words (ex as diagram)

yi	ge	ping-guo	'an apple'
1	MEAS	apple	

- Chinese aspect words

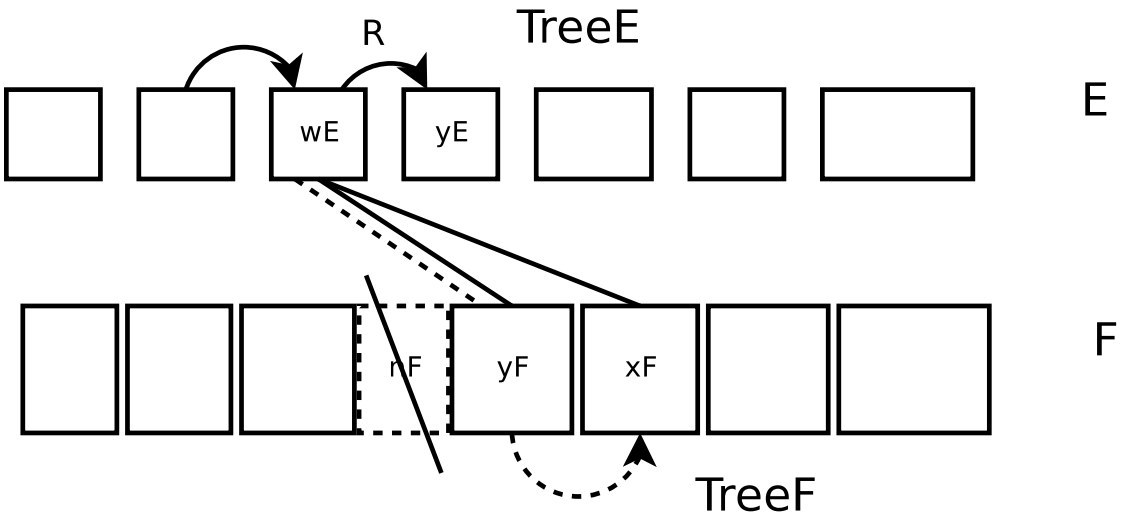
qu	le	'went' or 'to have gone'
go	COMPLETE	

these emerge as 1En-to-manyZh and unaligned-Zh cases.

Revised DPA: Revision 1: "head-initial"

revised 1-to-many rule

rather than creating n_f, just assume that the left-most F word is the head and draw dependencies from there.



Revised DPA: Revision 2: "Zh-side cleanup"

Restricted themselves to:

- closed class items
- POS info projected from En
- easily listed lexical categories

an example: if a series of Zh words are aligned with an En noun, make the rightmost word the head.

(Chinese is right-headed in nominal system, left-headed elsewhere)

Revised DPA: Revision 2: "Zh-side cleanup"

other examples include:

- enchain 'de' linking subordinator
- currency handling

(wanna look at the rules? They're in a tech report, so you'll have to write Dr. Hwa)

Results

Method	Precision	Recall	F-measure
DPA	34.5	42.5	38.1
RDPA 1 (head-initial)	59.4	59.4	59.4
RDPA 1+2 (h-i & rules)	68.0	66.6	67.3

total 76.6% F-measure gain over baseline(!)

Discussion

application of minimal linguistic knowledge to transfer information from one language to another

on the MT pyramid – low-middle approach, but much syntax gained!

potential applications for MT?

- learn syntactic relations from translations of well-parsed E
- learn phrase boundaries?

Stats MT (mostly) doesn't use DCA – how can these be combined?