Capturing Translational Divergences with a Statistical Tree-to-Tree Aligner

Mary Hearne, John Tinsley, Ventsislav Zhechev & Andy Way

National Centre for Language Technology
School of Computing
Dublin City University

Parallel treebanks

A parallel treebank comprises:

- sentence pairs
- parsed
- word-aligned
- tree-aligned

(Volk & Samuelsson, 2004)

The role of alignments:

Santos (1996), paraphrasing Lab (1990):

> *Having a linguistic description of two languages is not the same as having a linguistic description of the translation between them.*

Parallel treebanks

- Our work involves automatically obtaining a parallel treebank from a parallel corpus via *parsing* and *tree alignment*.
- Our overall objective is to use the parallel treebank for inducing a variety of syntax-aware and syntax-driven models of translation for use in data-driven MT.
- In this paper/presentation, the focus is on the capture of translational divergences through the application of a tree-aligner to gold-standard tree pairs.

Capturing translational divergences

We aim to:

- make explicit the syntactic divergences between source and target sentence pairs
- align to express as precisely as possible the translational equivalences between the tree pair
- constraining phrase-alignments in the data set is a consequence of aligning trees, but not an objective

We remain agnostic with regard to:

- which linguistic formalism is most appropriate for the expression of monolingual syntax
- how best to exploit parallel treebanks for syntax-aware data-driven MT

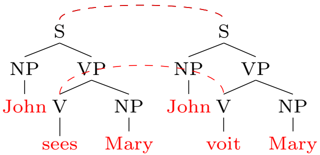
Outline

Outline

Tree-to-Tree Alignment

Links indicate *translational equivalence*:

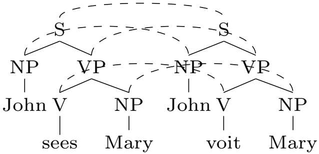
- a link between root nodes indicates equivalence between the sentence pair
- a link between any given pair of source and target nodes indicates
  - equivalence between the substrings they dominate
  - equivalence between the substrings they do *not* dominate



Tree-to-Tree Alignment

In the simplest case:

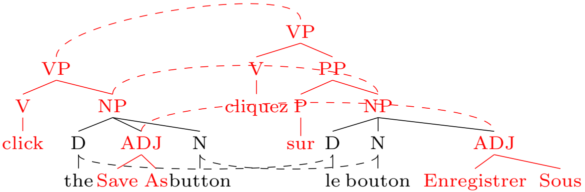
- the sentence lengths are identical
- the word order is identical
- the tree structures are isomorphic



Tree-to-Tree Alignment

Slightly more complex:

- not every node in each tree needs to be linked
- each node is linked at most once
- terminal nodes are not linked

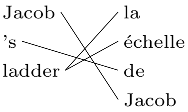


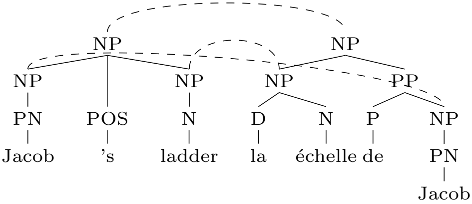
Tree-Alignment vs. Word-Alignment

Word-alignment: unaligned words are problematic and to be avoided
Tree-alignment: unaligned nodes are informative

... Jacob's ladder ... ⟶ ... l'échelle de Jacob ...

Word alignment: Tree alignment:





Hierarchical alignments

On the relationship between *'s* and *de* in

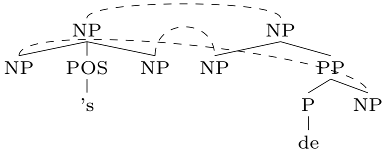
... Jacob's ladder ... ⟶ ... l'échelle de Jacob ...

$$\text{'}s \longrightarrow de$$

$$X\ \text{'}s\ Y \longrightarrow Y\ de\ X$$

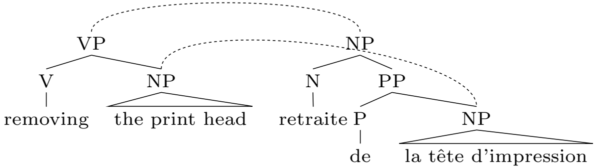
$$NP_1\ \text{'}s\ NP_2 \longrightarrow NP_2\ de\ NP_1$$

$$NP \rightarrow NP_1\ \text{'}s\ NP_2 \quad : \quad NP \rightarrow NP_2\ de\ NP_1$$

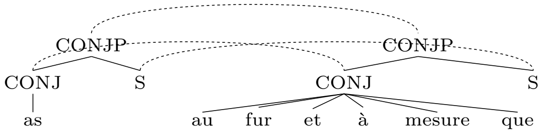


Outline

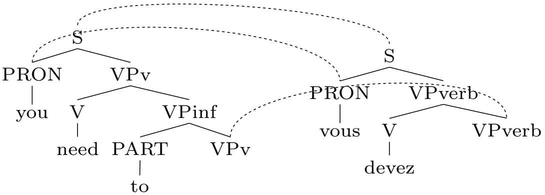
Nominalisation



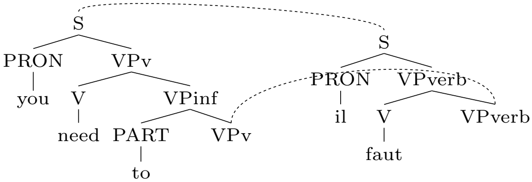
Lexical Divergences



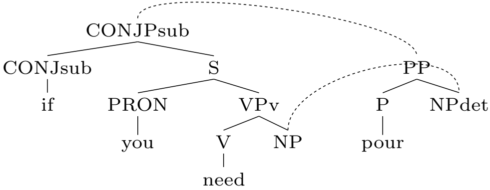
Context-Dependent Lexical Selection



(1)

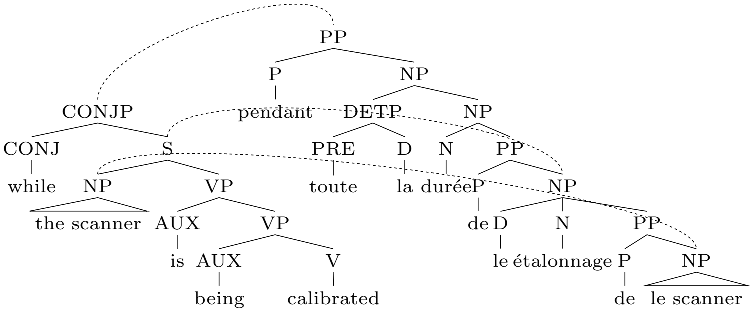


(2)

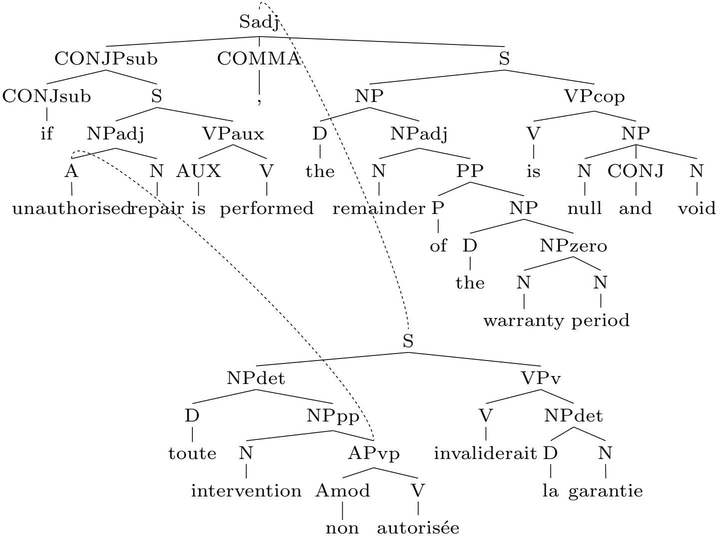


(3)

Embedded Complexities



Structural Dissimilarity



'any unauthorised action would invalidate the guarantee'

Outline

Tree-alignment algorithm

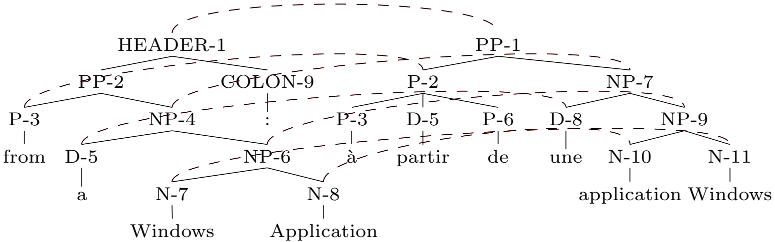
Alignment algorithm:

- hypothesise initial alignments: each source node can link to any target node and vice versa;
- assign a score to each hypothesised alignment;
- select a set of links meeting the well-formedness criteria according to a greedy search.

Well-formedness criteria:

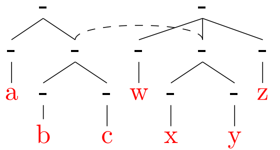
- each node can only be linked once;
- descendants of a source linked node may only link to descendents of its target linked counterpart;
- ancestors of a source linked node may only link to ancestors of its target linked counterpart.

Tree-alignment algorithm



	1	2	3	5	6	7	8	9	10	11
1	**1**	0	0	0	0	0	0	0	0	0
2	~~1~~	0	0	0	0	0	0	0	0	0
3	0	**3**	0	0	0	0	0	0	0	0
4	0	0	0	0	0	**6**	0	0	0	0
5	0	0	0	0	~~2~~	0	**2**	0	0	0
6	0	0	0	0	0	~~2~~	0	**5**	~~4~~	0
7	0	0	0	0	~~3~~	0	0	0	0	**7**
8	0	0	0	0	0	0	0	0	**4**	0
9	0	0	0	0	~~3~~	0	~~2~~	0	0	~~5~~

Tree-alignment algorithm



$s_l \;=\;$ b c
$t_l \;=\;$ x y
$\overline{s_l} \;=\;$ a
$\overline{t_l} \;=\;$ w z

Computing hypothesis scores:

Assume tree pair <S,T>, hypothesis <s,t>, the following strings and GIZA++ / Moses word-alignment probabilities.

$$s_l = s_i ... s_{ix} \quad \overline{s_l} = S_1 ... s_{i-1} s_{ix+1} ... S_m$$
$$t_l = t_j ... t_{jx} \quad \overline{t_l} = T_1 ... t_{j-1} t_{jx+1} ... T_n$$

Hypothesis score: $\gamma(\langle s,t \rangle) = \alpha(s_l|t_l)\ \alpha(t_l|s_l)\ \alpha(\overline{s_l}|\overline{t_l})\ \alpha(\overline{t_l}|\overline{s_l})$

String correspondence score: $\alpha(x|y) = \prod_{j=1}^{|x|} \frac{\sum_{i=1}^{|y|} P(x_j|y_i)}{|y|}$

Outline

Methodology

- dataset: HomeCentre English-French corpus, parsed and aligned, 810 sentence pairs
- Alignment evaluation:
  - precision and recall of automatic alignments vs. manual alignments
- Translation evaluation:
  - split the data into training and test, 6 splits, averaged results
  - MT system used: DOT (Hearne & Way, EAMT-06)
  - train the system on manual vs. automatic alignments
- Manual analysis of translational divergences

Alignment Evaluation vs. Gold Standard

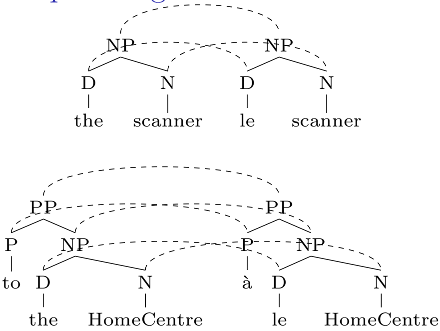
	Alignment Evaluation					
	all links		*lexical links*		*non-lexical links*	
Configs	Precision	Recall	Precision	Recall	Precision	Recall
scr1	0.6162	0.7783	0.5057	0.7441	**0.8394**	0.7486
scr2	0.6215	0.7876	0.5131	0.7431	0.8107	0.7756
scr1_sp1	**0.6256**	**0.8100**	0.5163	**0.7626**	0.8139	**0.8002**
scr2_sp1	0.6245	0.7962	**0.5184**	0.7517	0.8031	0.7871

Translation Evaluation vs. Gold Standard

	Translation Evaluation			
	(all links)			
Configs	Bleu	NIST	Meteor	Coverage
manual	0.5222	6.8931	71.8531	68.5417
scr1	0.5091	6.9145	71.7764	71.8750
scr2	**0.5333**	6.8855	**72.9614**	**72.5000**
scr1_sp1	0.5273	**6.9384**	72.7157	**72.5000**
scr2_sp1	0.5290	6.8762	72.8765	**72.5000**

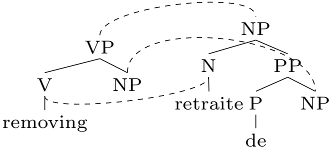
Capturing Translational Divergence

Simple, isomorphic alignments:



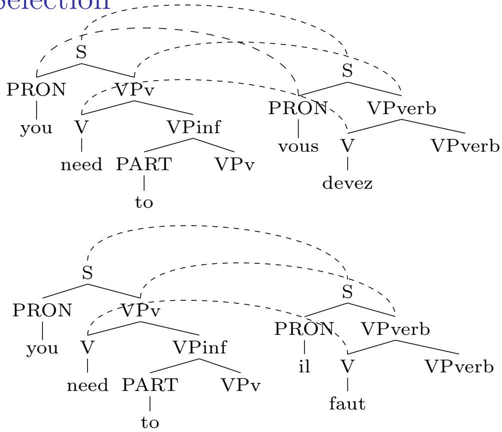
Capturing Translational Divergence

Nominalisation



Capturing Translational Divergence

Lexical Selection



Outline

Conclusions

- aligner performance is better at the phrase level than the lexical level
- imbalance between precision and recall at the lexical level
  - aligner uses GIZA++ word-alignment probabilities
  - GIZA++ prioritises broad coverage over high precision
  - in terms of capturing translational divergences between tree pairs, the preference is for the opposite

- it is appropriate for tree-alignment to prioritise precision over recall
- MT systems should use high-precision tree alignments *in conjunction with* broad-coverage models to preserve robustness

Future Work

- investigate alternative word-alignment methods to further improve the accuracy of the tree-alignment algorithm
- investigate the impact of imperfect parse quality on tree-alignment
- investigate the extraction of translation models from automatically-annotated parallel treebanks

The End.