Tree-based Alignment and Translation

John Tinsley, Ventsislav Zhechev, Mary Hearne and Andy Way

June 18, 2007

Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work

Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work

Parallel treebanks

A parallel treebank comprises:

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

(Volk & Samuelsson, 2004)

Parallel treebanks

Some areas of application for parallel treebanks are:

- training for data-driven machine translation
- knowledge source for transfer-rule induction
- reference for phrase-alignment
- knowledge source for corpus-based translation studies
- knowledge source for studies in contrastive linguistics

Automatic parallel treebank acquisition

Required elements:

- sentence pairs: increasing availability of bitexts, both manually and automatically constructed
- parsed: increasing availability of treebanks and trainable parsing resources
- word-aligned: availability of tools such as Giza++ and Moses
- tree-aligned: ???

Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

Future Work

Tree-alignment algorithm (Tinsley et al., 2007)

Principles:

- independence with respect to language pair and constituent labelling schema;
- preservation of the given tree structures;
- minimal external resources required;
- word-level alignments not fixed *a priori*.

Tree-alignment algorithm (Tinsley et al., 2007)

Well-formedness criteria:

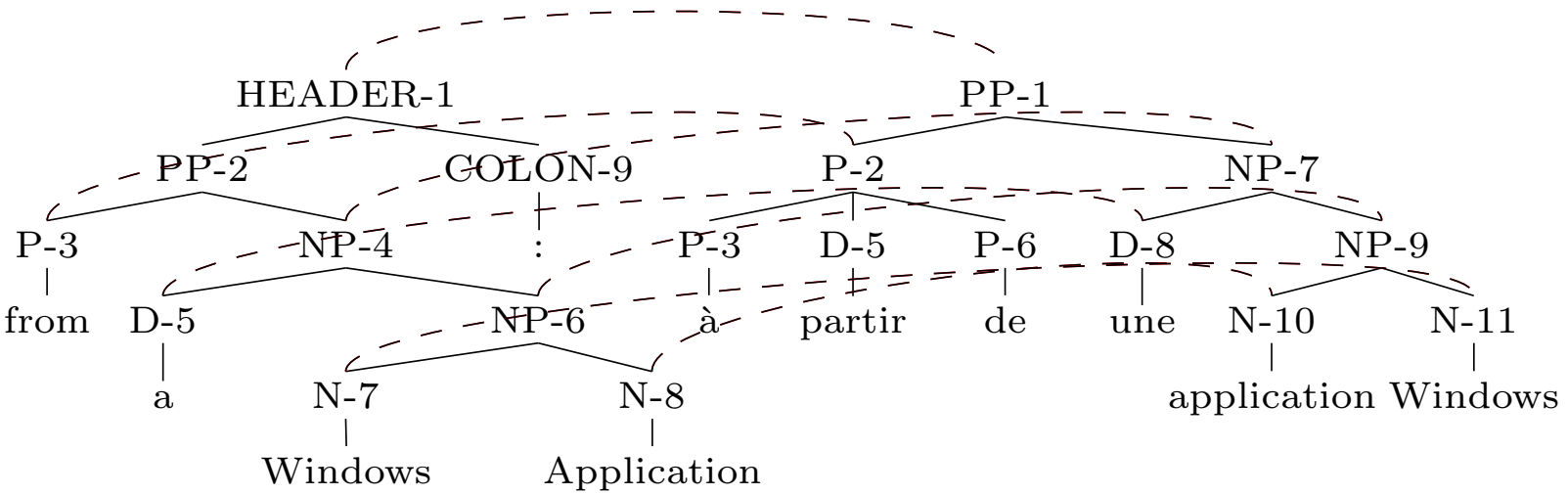
- a node can only be linked once;
- crossing constraints:
  - descendents of a source linked node may only link to descendents of its target linked counterpart;
  - ancestors of a source linked node may only link to ancestors of its target linked counterpart.

Tree-alignment algorithm (Tinsley et al., 2007)

Alignment algorithm:

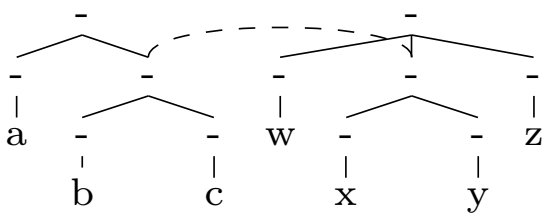
- hypothesise initial alignments: each source node can link to any target node and vice versa;
- assign a score to each hypothesised alignment;
- select a set of links meeting the well-formedness criteria according to a greedy search.

Tree-alignment algorithm (Tinsley et al., 2007)



	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 (Tinsley et al., 2007)



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.

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

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

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

Evaluation

	Alignment Evaluation					
	all links		*lexical links*		*non-lexical links*	
Configs	Precision	Recall	Precision	Recall	Precision	Recall
scr1	0.6686	0.7733	0.5615	0.7494	**0.8468**	0.7563
scr2	0.6736	0.7844	0.5722	0.7582	0.8163	0.7754
scr1_sp1	0.6763	**0.8026**	0.5733	**0.7675**	0.8191	**0.8020**
scr2_sp1	**0.6781**	0.7927	**0.5781**	0.7632	0.8140	0.7855

Evaluation

	Translation Evaluation			
	(all links)			
Configs	Bleu	NIST	Meteor	Coverage
manual	**0.5444**	**7.0701**	**0.7302**	70.4167
scr1	0.5163	6.8685	0.7242	73.7500
scr2	*0.5280*	*6.8869*	*0.7297*	73.5417
scr1_sp1	0.5252	6.8842	0.7285	***75.3125***
scr2_sp1	0.5258	6.8506	0.7268	73.6458

Outline

Parallel Treebanks

Automatic Tree-to-Tree Alignment

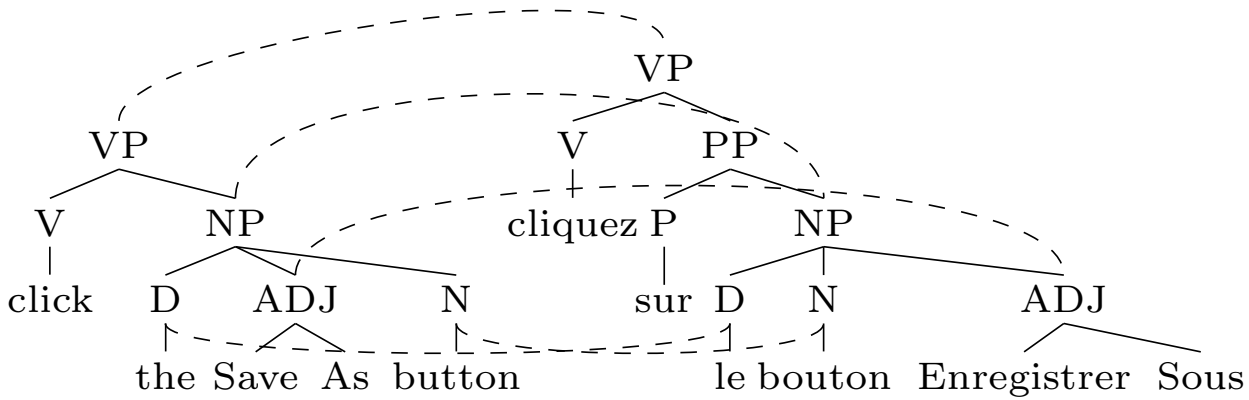
Future Work

Improving alignment quality

- alternative scoring strategies
- alternative search strategies
- using monolingual parse n-best lists for increased flexibility
- larger-scale translation-based evaluation

Using tree-alignments for MT

Aligner output:



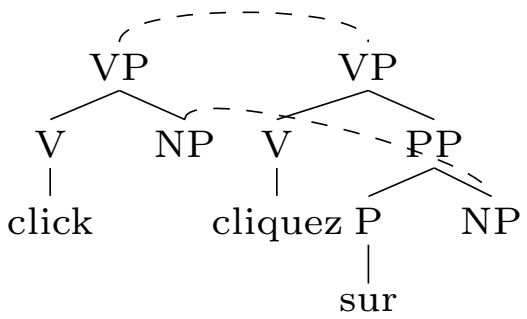
Resource extraction:

$$\textit{click} \longrightarrow \textit{cliquez sur}$$

$$\textit{click}\ \text{X} \longrightarrow \textit{cliquez sur}\ \text{X}$$

$$\textit{click}\ \text{NP} \longrightarrow \textit{cliquez sur}\ \text{NP}$$

$$\text{VP} \rightarrow \textit{click}\ \text{NP} \quad : \quad \text{VP} \rightarrow \textit{cliquez sur}\ \text{NP}$$



Generating and using alternative outputs

- generalise the existing aligner to the
  - string-to-tree
  - tree-to-string
  - string-to-string

  cases, and develop corresponding resource-extraction methods.