Cambridge ALTA
Institute for Automated Language Teaching and Assessment

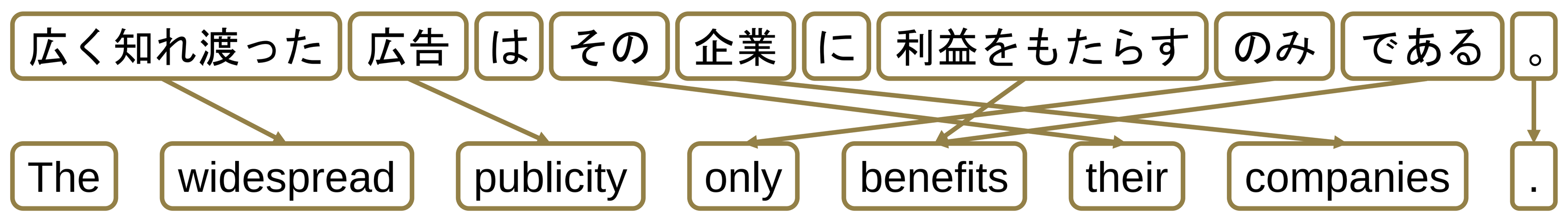
UNIVERSITY OF CAMBRIDGE

Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments

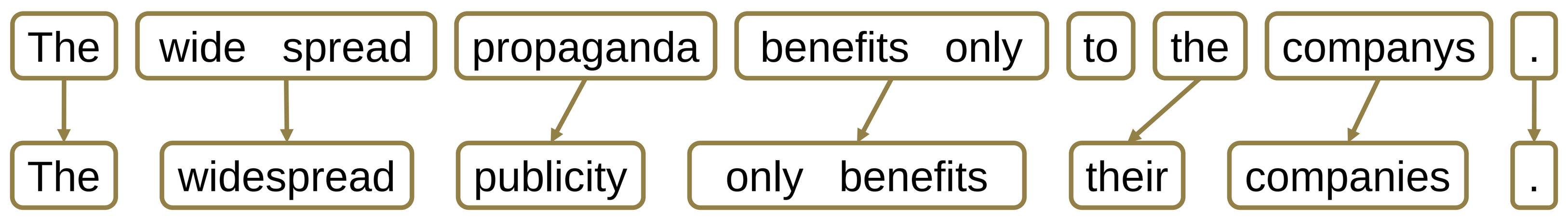
Mariano Felice, Christopher Bryant and Ted Briscoe

The Task: An Analogy

- Alignment in Machine Translation (MT)

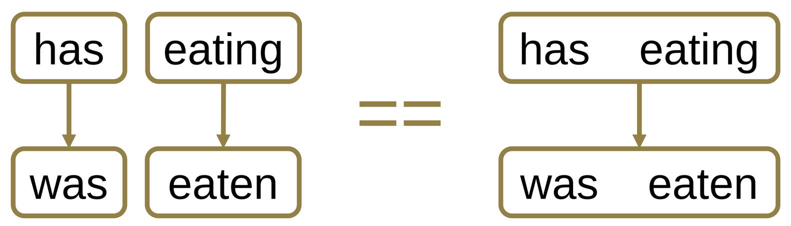


- Alignment in Grammatical Error Correction (GEC)



Observations

- Unlike in MT, most tokens in ESL sentences align with themselves.
- Error correction data is typically aligned manually or not at all.
- Human alignments can be inconsistent:

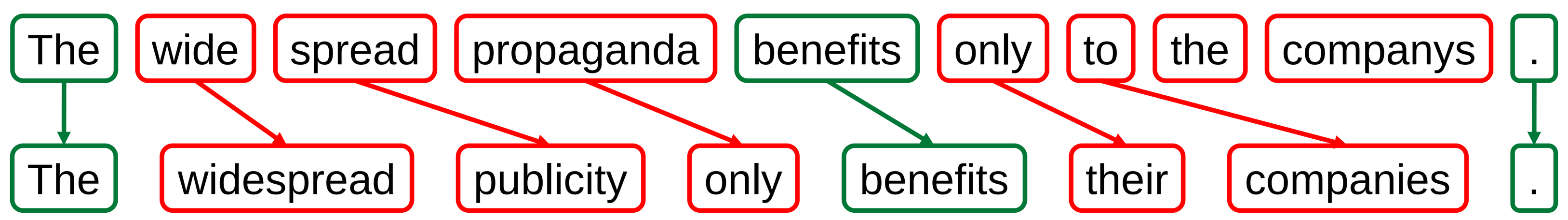


Goals

- Given parallel original and corrected sentences, extract edits automatically. The extracted edits should resemble human edits.
- Standardise treatment of ambiguous alignments in all datasets.
- Simplify the annotation of new data (reduce annotator burden).
- Facilitate more detailed evaluation of unannotated GEC system output.

Baseline: Levenshtein Alignment

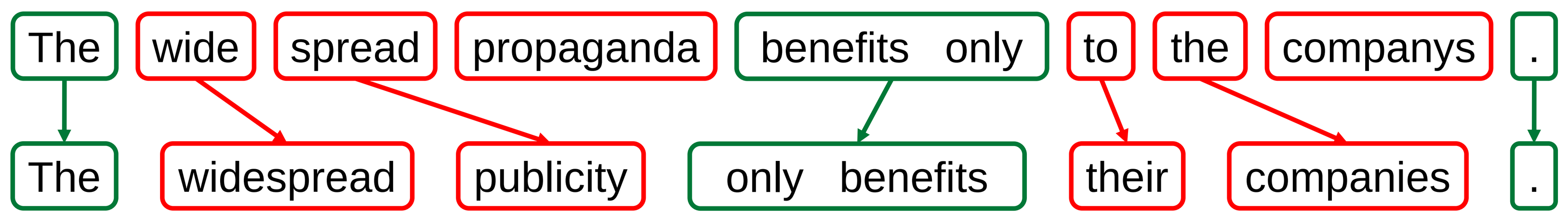
- Levenshtein finds a minimal way of transforming one sentence into another.



- Limitations:
  1. Word order errors are treated as insertions and deletions.
  2. Many alignments do not make linguistic sense.
  3. Cannot handle multi-token alignments.

Step 1: Damerau-Levenshtein Alignment

- An extension to Levenshtein that handles two-token “transpositions”.



- Further modified to handle transpositions of arbitrary length.
- Allows us to correctly extract word order errors.

Step 2: Linguistically Enhanced Damerau-Levenshtein

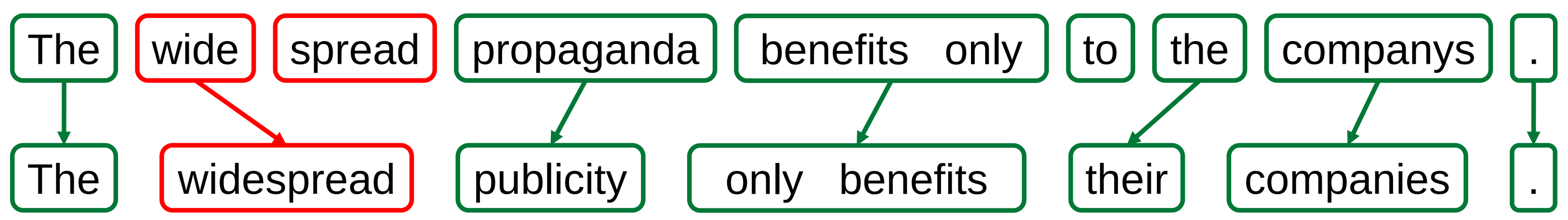
- Human alignments seem intuitively guided by linguistic knowledge.
- We therefore incorporated linguistic information into the token substitution cost of the automatic alignment:

$$\text{cost}_{\text{lemma}} = \begin{cases} 0 & \text{if same lemma,} \\ 0.499 & \text{otherwise} \end{cases}$$

$$\text{cost}_{\text{pos}} = \begin{cases} 0 & \text{if same pos,} \\ 0.25 & \text{if both content,} \\ 0.5 & \text{otherwise} \end{cases}$$

$$\text{cost}_{\text{char}} = \frac{\text{character alignment cost}}{\text{alignment length}}$$

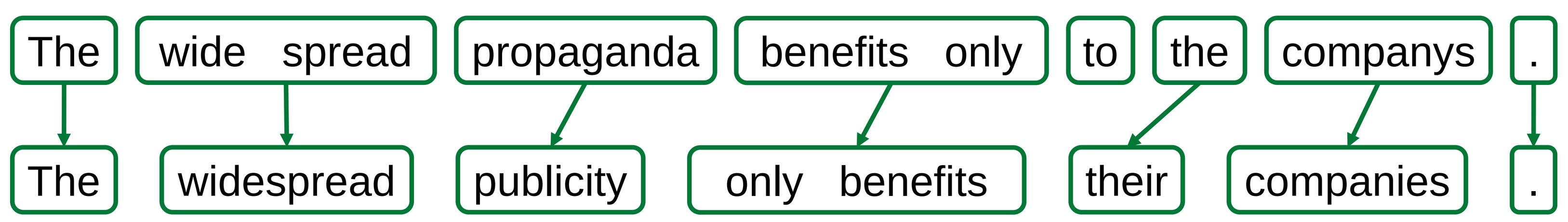
$$\text{cost}_{\text{substitution}} = \text{cost}_{\text{lemma}} + \text{cost}_{\text{pos}} + \text{cost}_{\text{char}}$$



- The added linguistic information helps produce more human-like alignments.

Step 3: Merging

- Multi-token edits account for only 20-30% of all edits in a typical dataset.
- More than half of these multi-token edits involve no more than two tokens.
- We wrote ten rules to handle the most salient patterns; e.g.
  - Whitespace errors: wide spread → widespread
  - Possessive errors: friends → friend 's
  - Phrasal Verbs: look at → watch



- Merging alignments enables us to capture multi-token edits.

Evaluation

- Compare human and auto edit spans in several publicly available datasets to compute an F-score.
- Human edit spans were minimised to remove unchanged tokens:
  - E.g. is eating → was eating == is → was
- Performance was measured in different settings:
  - All-split – Nothing is merged.
  - All-merge – All adjacent non-matches are merged.
  - This work – Rules decide whether adjacent non-matches are merged.

Dataset	Merging	Edit Extraction					
		TP	FP	FN	P	R	F_1
CoNLL 2013	All-split	2715	1612	659	62.75	80.47	70.51
	All-merge	2194	653	1180	77.06	65.03	70.54
	This work	2784	591	590	82.49	82.51	82.50
CoNLL 2014 (0)	All-split	1858	1320	526	58.46	77.94	66.81
	All-merge	1662	415	722	80.02	69.71	74.51
	This work	1893	550	491	77.49	79.40	78.43
CoNLL 2014 (1)	All-split	2635	1699	651	60.80	80.19	69.16
	All-merge	2435	554	851	81.47	74.10	77.61
	This work	2866	598	420	82.74	87.22	84.92
FCE-test	All-split	3660	1936	847	65.40	81.21	72.45
	All-merge	3144	778	1363	80.16	69.76	74.60
	This work	3861	739	646	83.93	85.67	84.79

- We also evaluated against previous approaches in an end-to-end system that uses a maximum entropy classifier to predict error types.
  - Swanson and Yamangil (2012) used a Levenshtein alignment and merged all adjacent non-matches.
  - Xue and Hwa (2014) used a Levenshtein alignment and maximum entropy merging classifier.

Dataset	Method	Edit Extraction F_1	Edit Extraction + Classification F_1
CoNLL 2013	S&Y	70.42	52.85
	X&H	74.07	55.89
	This work	82.50	61.40
CoNLL 2014 (0)	S&Y	72.92	46.95
	X&H	74.25	49.15
	This work	78.43	51.46
CoNLL 2014 (1)	S&Y	76.39	56.18
	X&H	79.21	59.24
	This work	84.92	63.38
FCE-test	S&Y	73.59	59.80
	X&H	79.18	65.33
	This work	84.79	69.88

- Our method outperforms all previous methods on all datasets.