Concept Alignment for Multilingual Machine Translation

04.07.2021

Arianna Masciolini

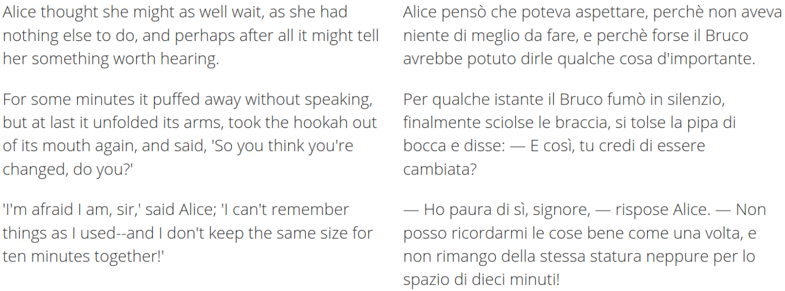
Context

- GF is well suited for domain-specific MT systems where precision is more important than coverage, as it provides strong guarantees of grammatical correctness
- in such systems, **lexical exactness** is as important as grammaticality
  - need for high-quality **translation lexica** preserving semantics *and* morphological correctness

The problem

- manually building a translation lexicon
  - is time consuming
  - requires significant linguistic knowledge
- desire to **automate** this process at least in part
  - possible when **example parallel data** are available

A parallel corpus



Alice thought she might as well wait, as she had nothing else to do, and perhaps after all it might tell her something worth hearing.	Alice pensò che poteva aspettare, perchè non aveva niente di meglio da fare, e perchè forse il Bruco avrebbe potuto dirle qualche cosa d'importante.
For some minutes it puffed away without speaking, but at last it unfolded its arms, took the hookah out of its mouth again, and said, 'So you think you're changed, do you?'	Per qualche istante il Bruco fumò in silenzio, finalmente sciolse le braccia, si tolse la pipa di bocca e disse: — E così, tu credi di essere cambiata?
'I'm afraid I am, sir,' said Alice; 'I can't remember things as I used--and I don't keep the same size for ten minutes together!'	— Ho paura di sì, signore, — rispose Alice. — Non posso ricordarmi le cose bene come una volta, e non rimango della stessa statura neppure per lo spazio di dieci minuti!

From Lewis Carroll, *Alice's adventures in Wonderland*. Parallel text at `paralleltext.io`

Alignment

Word alignment:

Alice thought she might as well wait, as she had nothing else to do, and perhaps after all it might tell her something worth hearing.

Alice pensò che poteva aspettare, perchè non aveva niente di meglio da fare, e perchè forse il Bruco avrebbe potuto dirle qualche cosa d'importante.

Phrase alignment:

Alice thought she might as well wait, as she had nothing else to do, and perhaps after all it might tell her something worth hearing.

Alice pensò che poteva aspettare, perchè non aveva niente di meglio da fare, e perchè forse il Bruco avrebbe potuto dirle qualche cosa d'importante.

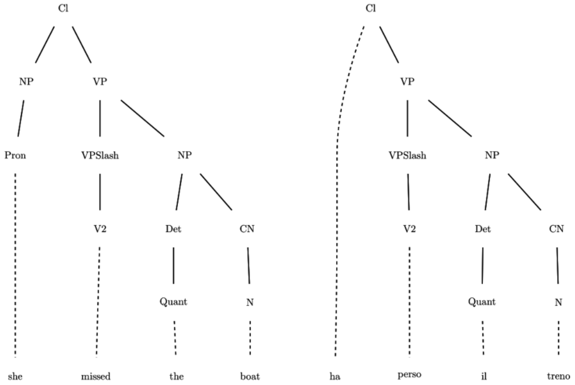
Statistical approaches

Standard approaches are statistical (IBM models).

- **Pros**:
  - easy to use
  - can handle noisy data
  - fast on large corpora
- **Cons**:
  - *require* large amounts of raw data
  - correspondences between strings → no morphological info
  - “fixed” level of abstraction (word or phrase)

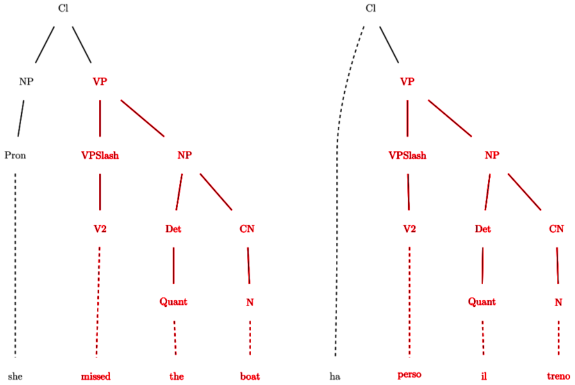
Syntax-based approaches I

Alternative: tree-to-tree alignment.



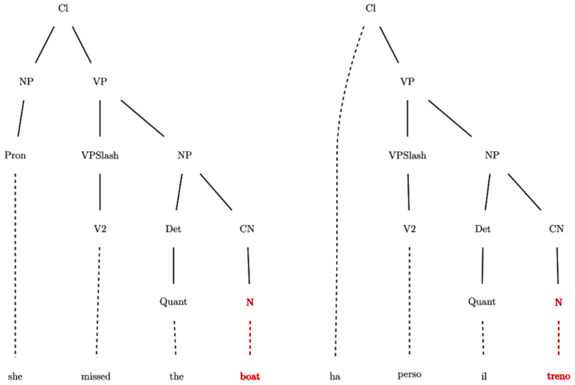
Syntax-based approaches II

Alternative: tree-to-tree alignment.



Syntax-based approaches III

Alternative: tree-to-tree alignment.



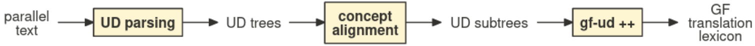
Comparison

statistical	**syntax-based**
require large amounts of raw data	work even on single *analyzed* sentence pairs
correspondences between strings	correspondences between grammatical objects
“fixed” level of abstraction	all levels of abstraction → **concept** alignment

Why not just use GF?

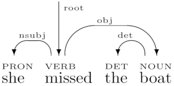
- quality of the analysis is crucial
  - lack of robust GF parsers
- dependency trees are an easier target for a parser
  - robust parsers such as UDPipe

Overview



1. parse parallel data to UD trees
2. search for aligned UD subtrees
3. convert them to GF trees and then grammar rules

UD trees



```
# text = she missed the boat
1 she she PRON _ _ 2 nsubj _ _
2 missed miss VERB _ _ 0 root _ _
3 the the DET _ _ 4 det _ _
4 boat boat NOUN _ _ 2 obj _
```

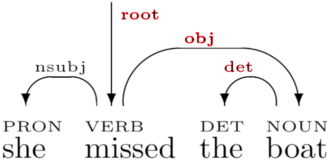
```
2 missed miss VERB _ _ 0 root _ _
  1 she she PRON _ _ 2 nsubj _ _
  4 boat boat NOUN _ _ 2 obj _
    3 the the DET _ _ 4 det _ _
```

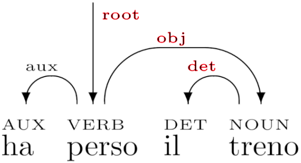
Graphical, CoNNL-U and Rose Tree representation of the same UD tree.

- dependency-labelled links between words (head-dependent pairs)
- POS tags
- . . .

Extracting concepts

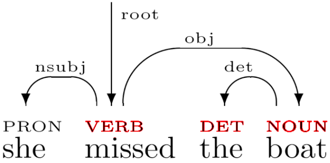
Matching dependency labels

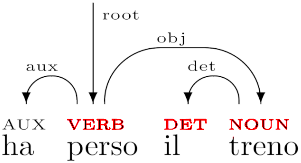




Aligning heads of maching trees

Using POS tags





Reusing known alignments

Translation divergences

Searching for specific patterns

Propagating concepts to a new language

Scenario 1

Scenario 2

Detailed overview

Generating grammar rules

Requirements

Morphological dictionaries

Extraction grammar

Lexical rules

Refining the generated lexicon

Interactive selection

Postprocessing

Conclusions

Summary

Questions?