Concept Alignment for Multilingual Machine Translation

04.07.2021

Arianna Masciolini

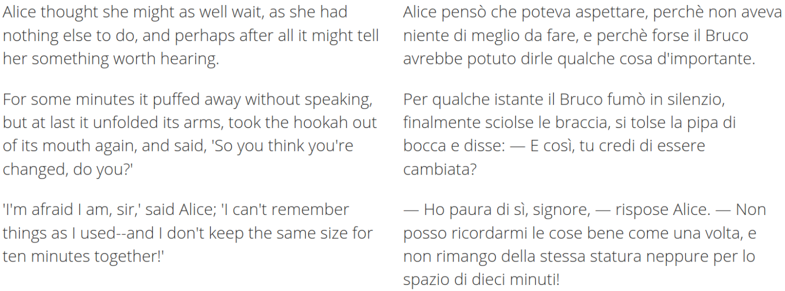
Context

- 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.

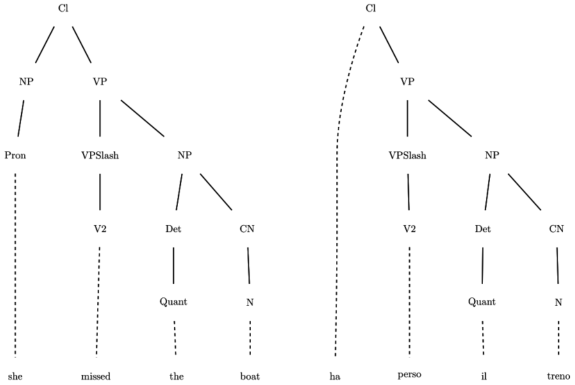
Statistical approaches

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)

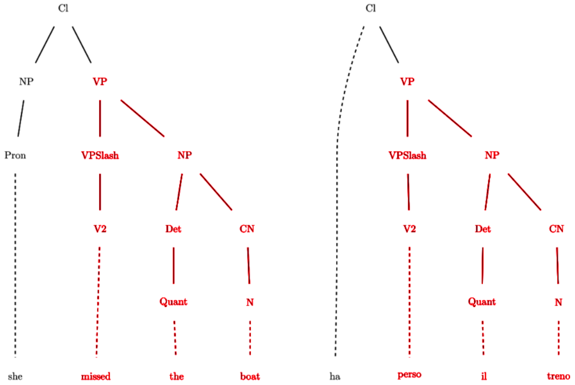
Syntax-based approaches I

Alternative: tree-to-tree alignment.



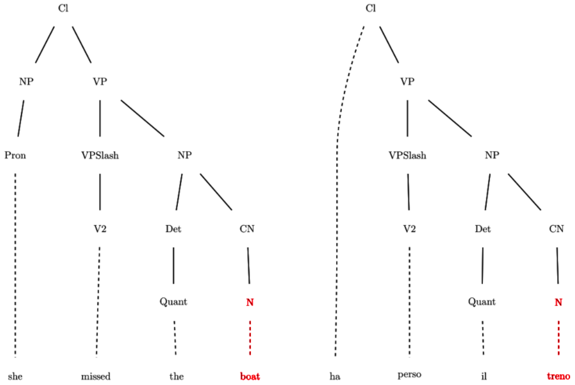
Syntax-based approaches II

Alternative: tree-to-tree alignment.



Syntax-based approaches III

Alternative: tree-to-tree alignment.



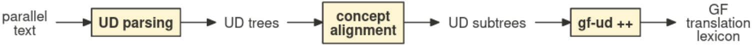
Comparison

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?

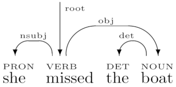
- quality of the analysis is crucial
    - lack of robust GF parsers
- dependency trees are an easier target for a parser
    - robust parsers such as UDPipe

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 _
```

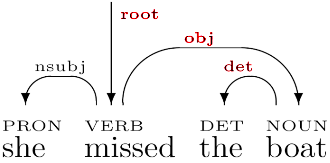
```
2 missed miss VERB _ _ 0 root _ _
 1 she she PRON _ _ 2 nsubj _ _
 4 boat boat NOUN _ _ 2 obj _
  3 the the DET _ _ 4 det _ _
```

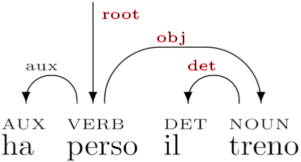
Graphical, CoNNL-U and Rose Tree representation of the same UD tree.

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

Extracting concepts

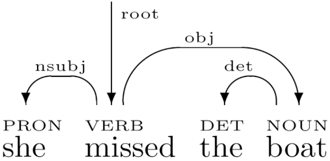
Matching dependency labels

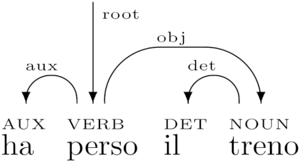




- ⟨*she missed the boat, ha perso il treno*⟩
- ⟨*missed the boat, perso il treno*⟩
- ⟨*the boat, il treno*⟩
- ⟨*the, il*⟩

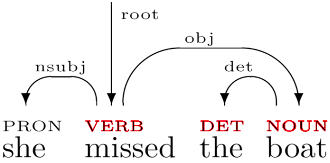
Aligning heads of maching trees

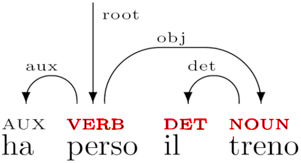




- ⟨*missed, ha perso*⟩
  - (incl. auxiliary in head)
- *⟨*boat, treno*⟩
- ⟨*the, il*⟩

Using POS tags





- more reliable ignoring function words
- in this case, same results as when matching labels
- can increase precision if used in conjuncion with labels
- can increase recall when labels do not coincide

Translation divergences

Divergence: systematic cross-linguistic distinction.

- categorial
  - ⟨*Gioara listens* ***distractedly****, Gioara lyssnar* ***distraherad***⟩
  - ⟨*Herbert completed his* ***doctoral*** *thesis, Herbert ha completato la sua tesi* ***di dottorato***⟩
- conflational
  - ⟨*Filippo is interested in* ***game development****, Filippo är intresserad av* ***spelutveckling***⟩
- structural
  - ⟨*I called* ***Francesco****, Ho telefonato* ***a Francesco***⟩
- head swapping
  - ⟨*Anna* ***usually*** *goes for walks, Anna* ***brukar*** *promenera*⟩
- thematic
  - ⟨***Yana*** *likes* ***books****,* ***A Yana*** *piacciono* ***i libri***⟩

Reusing known alignments

- Allows using CA in conjunction with statistical tools
- iterative application

Searching for specific patterns

- `gf-ud` pattern matching allows looking for specific syntactic patterns
- possible generalization via pattern replacement

Example predication patterns:

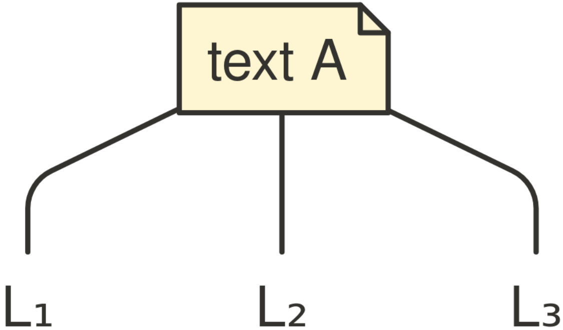
- ⟨*subj missed obj*,*subj ha perso obj*⟩
- ⟨*subj told iobj obj*,*subj berättade obj för obl*⟩

Propagating concepts to a new language

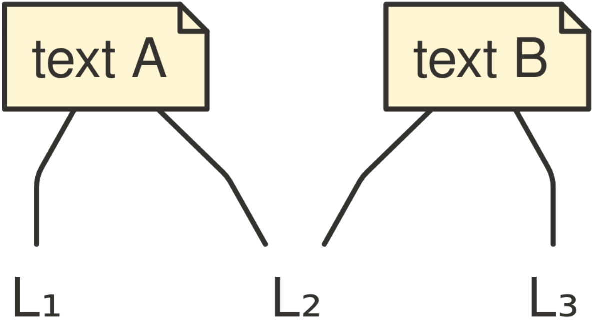
Concept Propagation

- So far, we focused on how to identify correspondences in bilingual parallel texts (*Concept Extraction*)
- what happens when we need to handle a third language?
    - *Concept Propagation*: finding expression corresponding to a known concept in a new language

Scenario 1



Scenario 2



Generating grammar rules

Requirements

- aligned UD trees
- extraction grammar
- morphological dictionaries

Morphological dictionaries

Purely morphological unilingual dictionaries.

Example:

```
...
lin morphologic_A =
  mkAMost "morphologic" "morphologicly" ;
lin morphological_A =
  mkAMost "morphological" "morphologically" ;
lin morphology_N =
  mkN "morphology" "morphologies" ;
...
```

Extraction grammar

Defines the syntactic categories and functions to build lexical entries.

Example (prepositional noun phrases):

```
PrepNP : Prep -> NP -> PP # case head
```

Lexical rules

Abstract:

```
fun in_the_field__inom_området_PP : PP ;
```

English concrete:

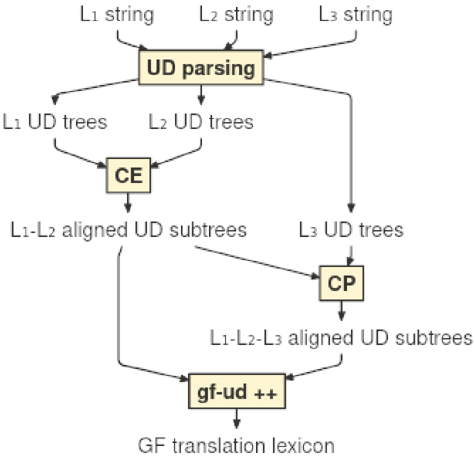
```
lin in_the_field__inom_område_PP =
  PrepNP in_Prep (DetCN the_Det (UseN field_N))
```

Refining the generated lexicon

Refining the generated lexicon

- interactive selection
- CoNNL-U synoptic viewer

Detailed overview



Summary

- concept extraction (UD)
- concept propagation (UD)
- GF lexicon generation
- postprocessing tools

Questions?