Grammar-Based Concept Alignment for Domain-Specific Machine Translation

08.09.2021

Arianna Masciolini and Aarne Ranta

Context

- In **domain-specific MT** precision is often more important than coverage
- grammar-based pipelines (cf. GF) provide strong guarantees of **grammatical correctness**
- **lexical exactness** is as important as grammaticality
  - need for high-quality **translation lexica** preserving semantics *and* morphological correctness

Translation lexica

- often built **manually**
  - **time** consuming
  - significant **linguistic knowledge** required
- need for at least partial **automation**
  - **example parallel data** required

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.

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?'

'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!'

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.

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?

— 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.

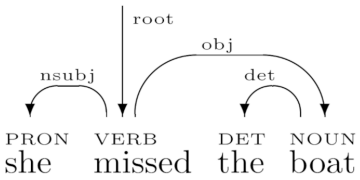
Approaches to automation

statistical (e.g. IBM models)	**syntax-based**
require **large amounts of data**	work consistently well even on **individual sentence pairs**
works with **raw** data	requires the data to be **analyzed**
correspondences between **strings**	correspondences between **grammatical objects**
"fixed" level of abstraction (**word** or **phrase**)	all levels of abstraction $\rightarrow$ **concept** alignment

Our approach

- inconsistencies between different grammar formalisms → translation lexicon implemented in **GF**
- lack of robust constituency parsers while high-quality analysis is crucial → **UD** parsing
- `gf-ud` for conversion

Universal Dependencies




```
# 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.

- framework for cross-linguistically consistent grammatical annotation
- dependency-labelled links between words (head-dependent pairs)
- cannot be used for target language generation

Grammatical Framework

- constituency grammar formalism/programming language for **multilingual grammars**
  - one abstract syntax
  - multiple concrete syntaxes
- compilation-like approach to translation (parsing + linearization)

Concept Extraction

Definitions

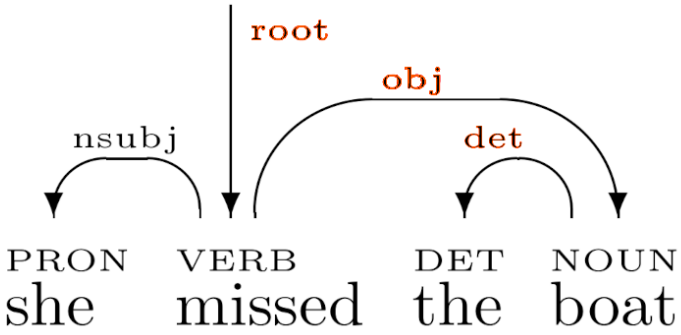
Concept: semantic unit of compositional translation expressed by a word or construction, conceived as a lemma equipped with morphological variations.

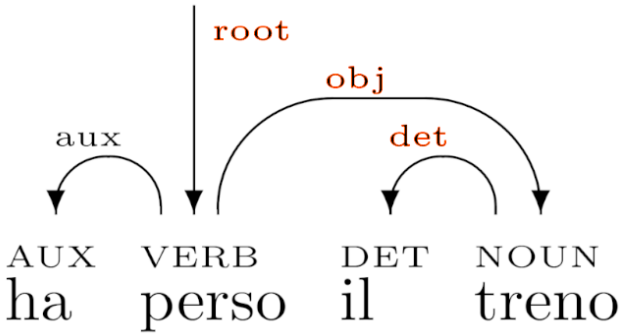
Alignment: tuple of equivalent concrete expressions in different languages; represents a concept.

Extraction algorithm

procedure EXTRACT($criteria$,(t, u))
    $alignments = \emptyset$
    if (t, u) matches any alignment $criteria$ **then**
        $alignments$ += (t, u)
        for (t', u') in SORT(SUBTS(t)) $\times$ SORT(SUBTS(u))
do
            extract($criteria$,(t', u'))
    return $alignments$

Matching UD labels



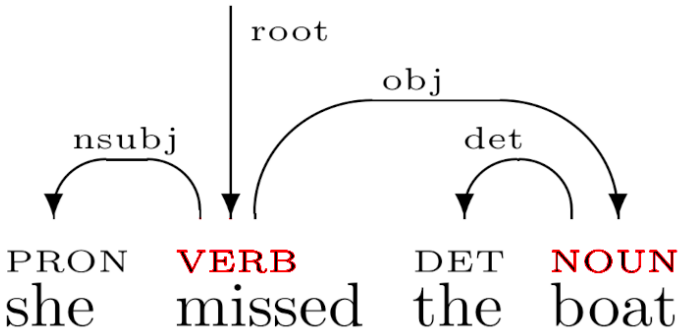


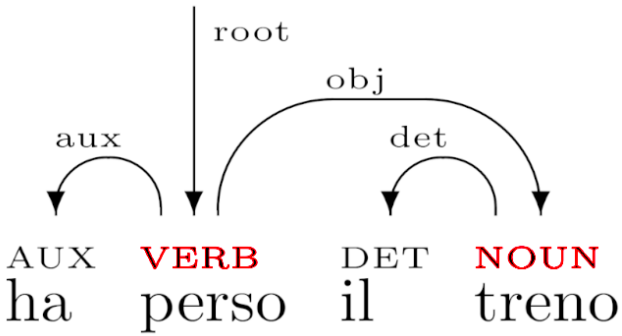
- $\langle$*she missed the boat, ha perso il treno*$\rangle$
- $\langle$*missed the boat, perso il treno*$\rangle$
- *$\langle$*the boat, il treno*$\rangle$
- $\langle$*the, il*$\rangle$

Simple improvement: aligning heads of matching subtrees

- $\langle$*she missed the boat, ha perso il treno*$\rangle$, $\langle$*missed the boat, perso il treno*$\rangle$ $\rightarrow$ $\langle$*missed, ha perso*$\rangle$ (including the auxiliary)
- $\langle$*the boat, il treno*$\rangle$ $\rightarrow$ *$\langle$*boat, treno*$\rangle$

POS equivalence





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

Known translation divergence

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***⟩

Known alignment

- 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:

- ⟨*she missed the boat, ha perso il treno*⟩ → ⟨*[subj] missed [obj], ha perso [obj]*⟩
- ⟨*she told you that, hon berättade det för dig*⟩ → ⟨*[subj] told [iobj] [obj],[subj] berättade [obj] för [obl]*⟩

Grammar rules generation

Requirements

- aligned UD trees
- `gf-ud`
- **morphological dictionaries**
- **extraction grammar**

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 NPs):

```
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))
```

Evaluation

Evaluating extraction

UD tree alignments are evaluated:

- independently from the quality of UD parsing (100-sentence subset of the manually annotated PUD corpus)
- on raw text (DMI and CSE course plans corpora)

Results on PUD corpus

	CE		`fast_align` (100 sentences)		`fast_align` (full dataset)	
	en-it	**en-sv**	**en-it**	**en-sv**	**en-it**	**en-sv**
distinct alignments	536	638	1242	1044	1286	1065
correct	392 (73%)	514 (80%)	346 (28%)	538 (52%)	540 (42%)	677 (64%)
usable in MT	363 (68%)	503 (79%)	316 (25%)	525 (50%)	510 (40%)	666 (63%)

- CE module compared with `fast_align`, so extracting only one-to-many and many-to-one alignments
- CE has much higher precision, even when `fast_align` is trained on full 1000-sentence corpus

Results on course plans corpora

	PUD (100 sentences)		course plans	
	en-it	en-sv	DMI (881 sentences)	CSE (539 sentences)
distinct alignments	1197	1325	1823	1950
correct	916 (77%)	1112 (85%)	1205 (66%)	1269 (66%)
usable in MT	880 (74%)	1099 (84%)	1157 (63%)	1248 (64%)

- comparison between experiments on manually annotated treebanks and raw text
- precision decreases, but is still higher than `fast_align`'s
- recall much lower

MT experiments

- no need to write an *ad hoc* grammar: extend extraction grammar with existing RGL functions
- 2 bilingual lexica from course plans corpora
- corpus of sentences to translate generated in the GF shell
  - semi-random lexical and grammatical variations on a set of semantically plausible sentences
- metric: BLEU scores
- reference translations obtained by manual postprocessing of the automatic ones
  - avoid low scores due to different but equally valid lexical and grammatical choices

Results

	DMI (en-it)	CSE (en-sv)
BLEU-1 to 4	55	61
BLEU-1 to 3	63	68
BLEU-1 to 2	70	74
BLEU-1	79	81

- better results for English-Swedish (due to systematic errors in Italian)
- sentence-level scores range from 0 (sometimes due to a single semantic error) to 100

Conclusions

- extraction technique performing consistently well on small datasets
- simultaneous extraction of word, phrase, ... alignments, incl. discontinuous expressions
- can be used to look for specific types of correspondences, e.g. predication patterns
- automatic generation of compilable, morphology-aware GF translation lexica
- configurable to handle language pair- or corpus-specific divergences
- requires manual corrections and completions, but can reduce the time required for bootstrapping translation lexica significantly
- consists of a Haskell library + executables

Current and future work

- Concept Propagation
  - same text in new language (equivalent to multilingual CE)
  - new text in new language (within same domain)
- integration with statistical tools
- postprocessing tools