Grammar-Based Concept Alignment for Domain-Specific Machine Translation

08.09.2021

Arianna Masciolini and Aarne Ranta

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

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

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

The problem

- manually building a translation lexicon
    - is time consuming
    - requires significant linguistic knowledge
- desire to **automate** this process at least in part

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

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

Alignment

Word alignment:

Alignment

Word alignment:

Phrase alignment:

Statistical approaches

Standard approaches are statistical (IBM models).

Statistical approaches

Standard approaches are statistical (IBM models).

- **pros**:

Statistical approaches

Standard approaches are statistical (IBM models).

- **pros**:
    - easy to use

Statistical approaches

Standard approaches are statistical (IBM models).

- **pros**:
  - easy to use
  - can handle noisy data

Statistical approaches

Standard approaches are statistical (IBM models).

- **pros**:
  - easy to use
  - can handle noisy data
  - fast on large corpora

Statistical approaches

Standard approaches are statistical (IBM models).

- **pros**:
  - easy to use
  - can handle noisy data
  - fast on large corpora
- **cons**:

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

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

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

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, phrase or sentence)

Grammar-based approaches

TODO: - relation to prev syntax-based work

Our approach

TODO: main features and advantages

Concept Alignment

TODO: definitions: - concepts: . . . - alignment: . . .

Grammatical Framework

- formalism/programming language to write **multilingual grammars** → solves problem 1
  - one abstract syntax
  - multiple concrete syntaxes

Grammatical Framework

- formalism/programming language to write **multilingual grammars** → solves problem 1
  - one abstract syntax
  - multiple concrete syntaxes
- compilation-like approach to translation → good, grammaticality-preserving target language generation
- but: problem 2 persist

Universal Dependencies

- framework for cross-linguistically consistent grammatical annotation → same “multilingual” approach as GF

Universal Dependencies

- framework for cross-linguistically consistent grammatical annotation → same “multilingual” approach as GF
- based on *dependency*, as opposed to constituency, relation
  - **dependency**: word-to-word correspondence
    - head
    - dependent in some relation with the head

Universal Dependencies

- framework for cross-linguistically consistent grammatical annotation → same “multilingual” approach as GF
- based on *dependency*, as opposed to constituency, relation
  - **dependency**: word-to-word correspondence
    - head
    - dependent in some relation with the head
- easier target for a parser (e.g. UDPipe) → solves problem 2
- but: cannot be used for target language generation

Solution: UD + GF

Concept Extraction

Extraction algorithm

Aligning heads of maching trees

- $\langle$*the boat, il treno*$\rangle$

Aligning heads of maching trees

- $\langle$*the boat, il treno*$\rangle \rightarrow$ *$\langle$*boat, treno*$\rangle$

Aligning heads of maching trees

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

Aligning heads of maching trees

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

Aligning heads of maching trees

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

Alignment criteria

TODO: list them all

Matching UD labels

Matching UD labels

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

Matching UD labels

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

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*$\rangle$

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*$\rangle$

POS equivalence

POS equivalence

- more reliable **ignoring function words**

POS equivalence

- more reliable **ignoring function words**
- in this case, basically same results as when matching labels

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

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.

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***⟩

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***⟩

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***⟩

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*⟩

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

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

Searching for specific patterns

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

Searching for specific patterns

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

Example predication patterns:

- $\langle$*she missed the boat, ha perso il treno*$\rangle \rightarrow \langle$*[$subj$] missed [$obj$], ha perso [$obj$]*$\rangle$
- $\langle$*she told you that, hon berättade det för dig*$\rangle \rightarrow \langle$*[$subj$] told [$iobj$] [$obj$],[$subj$] berättade [$obj$] för [$obl$]*$\rangle$

Grammar rules generation (TODO: shorten!)

Requirements

- aligned UD trees

Requirements

- aligned UD trees
- dependency configurayions for `gf-ud`

Requirements

- aligned UD trees
- dependency configurayions for `gf-ud`
- **morphological dictionaries**

Requirements

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

Morphological dictionaries

Purely morphological unilingual 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.

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 ;
```

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

Data

Evaluating extraction

TODO: strategy

Results on manually annotated tree-

TODO: table 1

Results on raw text

TODO: table 2

MT experiments

TODO: strategy

Results

TODO: tables 3-4or just 3 + comments

Conclusions

Future work

TODO: ? - [ ] a - [x] b