Syntax-based Concept Alignment for Machine Translation

Master's thesis, A.Y. 2020-2021

Arianna Masciolini

supervisor: Aarne Ranta
examiner: Carl-Johan Seger

Concept Alignment

A first definition

Concept Alignment: the task of finding semantical correspondences between parts of multilingual parallel texts.

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`

CA at different levels of abstraction

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.

CA at different levels of abstraction

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.

Subtasks

- **Concept Extraction**: identifying new concepts via linguistic comparison
- **Concept Propagation**: finding expressions corresponding to known concepts in a particular language

CA in translation

A human translator

1. recognizes concepts in the text to translate
2. looks for ways to render them in the target language

CA in translation

A human translator

1. recognizes concepts in the text to translate
2. looks for ways to render them in the target language

. . . same idea behind *compositional* Machine Translation.

Semantic compositionality

The meaning of a complex expression is determined by:

- the meanings of its components (lexical semantics)
- the way its components are combined with each other (syntax)

Semantic compositionality

The meaning of a complex expression is determined by:

- the meanings of its components (lexical semantics)
- the way its components are combined with each other (syntax)

The *translation* of a complex expression is given by:

- the *translations* of its components (lexical semantics)
- the way its components are combined with each other (syntax, taking cross-lingual divergences into account)

Statistical approaches

Standard approaches to automation are statistical (IBM models)

Issues:

- "fixed" level of abstraction (generally either word or phrase alignment)
- correspondences are between strings
- need large amounts of raw data

Syntax-based approaches

Alternative: tree-to-tree alignment, generally based on constituency grammars.

MISSING FIGURE (parse trees + micorgrammar)

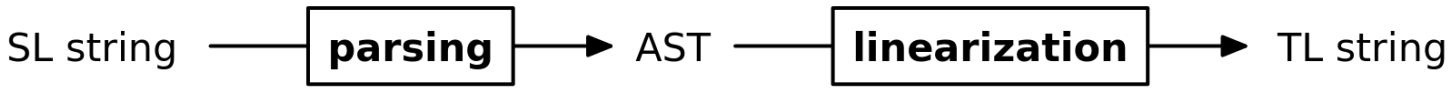
- ~~"fixed" level of abstraction~~ work at all levels of abstraction simultaneously
- correspondences are between ~~strings~~ grammatical objects
- ~~need large amounts of raw data~~ work consistently well even on single *analyzed* sentence pairs

Syntax-based approaches: issues

1. grammars often defined independently, so not compatible each other
2. lack of robust parsers, while the quality of the analyses is crucial

Grammatical Framework

- formalism/programming language to write **multilingual grammars** $\rightarrow$ solves problem 1
  - one abstract syntax
  - multiple concrete syntaxes
- compilation-like approach to translation $\rightarrow$ good, grammaticality-preserving target language generation



- but: problem 2 persist

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 MISSING FIGURE same parse tree as before (EN) + corresponding UD
- easier target for a parser (e.g. UDPipe) → solves problem 2
- but: cannot be used for target language generation

Solution: UD + GF