

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.

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

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

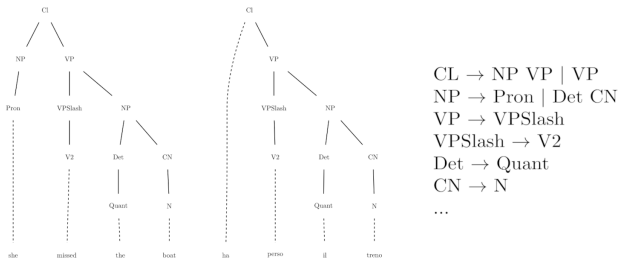
Standard approaches to automation are statistical (IBM models)

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.



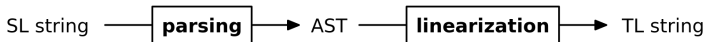
- ❑ ~~“fixed” level of abstraction~~ work at all levels of abstraction
- ❑ correspondences are between ~~strings~~ grammatical objects
- ❑ ~~need large amounts of raw data~~ work consistently well even on single *analyzed* sentence pairs

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

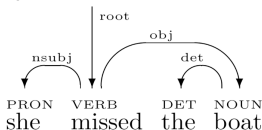
- ❖ formalism/programming language to write **multilingual grammars** → solves problem 1
 - ❖ one abstract syntax
 - ❖ multiple concrete syntaxes
- ❖ compilation-like approach to translation → good, grammaticality-preserving target language generation



- ❖ but: problem 2 persist

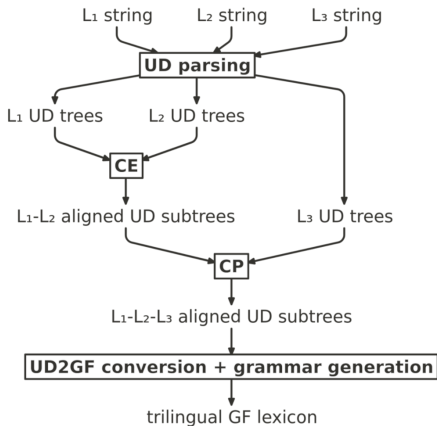
Universal Dependencies

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



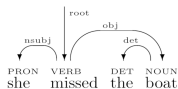
- ❖ easier target for a parser (e.g. UDPipe) → solves problem 2
- ❖ but: cannot be used for target language generation

Solution: UD + GF



Concept Extraction

Representations of 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 _ _

- ❖ CoNNL-U is the standard format for UD trees
- ❖ internally to the CA module, they are represented as rose trees

```
data RTree n = RTree n [RTree n]
```

```
type UDTree = RTree UDWord
```

```
type Alignment = (UDTree,UDTree)
```

- ❖ UDWord represents a line of a CoNNL-u file
- ❖ alignments are pairs of ud trees

Baseline

| | |
|-----------------------------------|-------------------------------------|
| 2 missed miss VERB _ _ 0 root _ _ | 2 perso perdere VERB _ _ 0 root _ _ |
| 1 she she PRON _ _ 2 nsubj _ _ | 1 ha avere AUX _ _ 2 aux _ _ |
| 4 boat boat NOUN _ _ 2 obj _ | 4 treno treno NOUN _ _ 2 obj _ _ |
| 3 the the DET _ _ 4 det _ _ | 3 il il DET _ _ 4 det _ _ |

1. recursively sort trees based on the UD label of their root node (not needed in this case)
2. pad the trees → perfectly aligned trees

| | |
|---|-------------------------------------|
| 2 missed miss VERB _ _ 0 root _ _ (dummy node replacing the aux) | 2 perso perdere VERB _ _ 0 root _ _ |
| 1 she she PRON _ _ 2 nsubj _ _ | 1 ha avere AUX _ _ 2 aux _ _ |
| 4 boat boat NOUN _ _ 2 obj _ | (dummy node replacing the nsubj) |
| 3 the the DET _ _ 4 det _ _ | 4 treno treno NOUN _ _ 2 obj _ _ |
| | 3 il il DET _ _ 4 det _ _ |

3. extract alignments:
 - ❖ subtrees: $\langle she missed the boat, ha perso il treno \rangle$, $\langle the boat, il treno \rangle$, $\langle the, il \rangle$
 - ❖ heads: $\langle missed, perso \rangle$, $\langle boat, treno \rangle$

Proposed improvements

- ❖ multiple alignment criteria → no perfect alignment
 - ❖ label matching
 - ❖ POS-equivalence
 - ❖ known alignment
 - ❖ known divergence
- ❖ better head alignment

Basic criteria

- ❖ **label matching** (original criterion): trees in matching context are aligned if they have the same UD label
- ❖ **POS-equivalence**: trees in matching context are aligned if they have the same multiset of POS tags of their *meaning-carrying* words
 - ❖ meaning-carrying words \simeq content words
- ❖ **known alignment**: trees in matching context are aligned if an equivalent alignment is already known
 - ❖ counting

Divergences

Divergence: systematic cross-linguistic distinctions.

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

Head alignment

- ❖ extremely useful when alignment is perfect, like $\langle \textit{Claudio eats a banana}, \textit{Claudio mangia una banana} \rangle$
 - ❖ $\langle \textit{eats}, \textit{mangia} \rangle$
 - ❖ $\langle \textit{banana}, \textit{banana} \rangle$
- ❖ many problematic cases
 - ❖ some types of divergences \rightarrow do not always align heads
 - ❖ compounds & head verbs with auxiliaries \rightarrow enhanced head alignment
 - $\langle \textit{many decisions were taken by Tommaso}, \textit{många viktiga beslut togs av Tommaso} \rangle \rightarrow \langle \textit{were taken}, \textit{togs} \rangle$
 - $\langle \textit{Giorgio took a course on machine learning techniques}, \textit{Giorgio deltog i en kurs om maskininlärningstekniker} \rangle \rightarrow \langle \textit{machine learning techniques}, \textit{maskininlärningstekniker} \rangle$

Evaluation on PUD treebanks

Against the baseline

| | baseline | | improved version | |
|------------------------|--------------|--------------|------------------|---------------|
| | en-it | en-sv | en-it | en-sv |
| distinct alignments | 1097 | 1257 | 1198 | 1314 |
| correct (+ and =) | 830 (58.12%) | 995 (79.15%) | 964 (80.46%) | 1105 (84.03%) |
| correct and useful (+) | 776 (54.34%) | 976 (77.64%) | 896 (74.79%) | 1082 (82.28%) |

Evaluation on PUD treebanks

Against the baseline

| | baseline | | improved version | |
|------------------------|--------------|--------------|------------------|---------------|
| | en-it | en-sv | en-it | en-sv |
| distinct alignments | 1097 | 1257 | 1198 | 1314 |
| correct (+ and =) | 830 (58.12%) | 995 (79.15%) | 964 (80.46%) | 1105 (84.03%) |
| correct and useful (+) | 776 (54.34%) | 976 (77.64%) | 896 (74.79%) | 1082 (82.28%) |

Against fast_align

| | our system | fast_align 100 | fast_align 1000 |
|---------------------|--------------|----------------|-----------------|
| distinct alignments | 716 | 1440 | 1435 |
| correct | 536 (74.86%) | 410 (28.47%) | 656 (45.71%) |
| correct and useful | 491 (68.57%) | 371 (25.76%) | 590 (41.11%) |

Evaluation on “raw” data

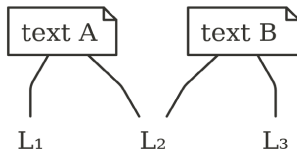
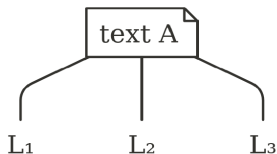
Data: sentence-aligned Computer Science course plans

- ❖ CSE (GU/Chalmers)
- ❖ DMI (UniPG)

| | DMI (en-it, 798 sentences) | CSE (en-sv, 498 sentences) |
|-------------------------------|----------------------------|----------------------------|
| distinct alignments | 352 | 529 |
| correct (+ and =) | 243 (69.03%) | 368 (69.56%) |
| correct and useful (+) | 229 (65.05%) | 351 (66.35%) |

Concept Propagation

Two scenarios



1. 3+ lingual parallel text
2. 2 bilingual parallel texts with one language in common

General algorithm

For each L_1 - L_2 alignment:

1. look for its L_2 member among all subtrees of the L_2 version of the text where it is to be propagated
2. if it is present, align the sentence it belongs to with its TL counterpart with the same procedure used for CE
3. if multiple candidate alignments are found, select the one with the closest depths

Caveats

- ❖ in step 1, irrelevant details of UD trees are to be ignored
 - ❖ only consider word form, lemma, POS tag and dependency relation
- ❖ head alignments require special treatment as they are not composed of subtrees

Evaluation: scenario 1

| | en-sv | it-sv |
|----------------------|---------------|--------------|
| propagated | 1019 (85.05%) | 979 (84.64%) |
| tot. errors | 133 (13.05%) | 187 (19.1%) |
| CP-introduced | 75 (56.39%) | 84 (44.91%) |

- ❖ PUD treebanks
- ❖ the vast majority of concepts is propagated

Evaluation: scenario 2

Texts in different domains (subsets of PUD treebanks)

| | en-it-sv | it-en-sv | en-sv-it | sv-en-it | it-sv-en | sv-it-en |
|----------------------|-------------|-------------|-------------|-------------|--------------|-------------|
| extracted | 638 | 638 | 687 | 687 | 608 | 608 |
| propagated | 92 (14.42%) | 92 (14.42%) | 98 (14.26%) | 84 (12.22%) | 101 (16.61%) | 87 (14.37%) |
| tot. errors | 46 (50%) | 21 (22.82%) | 42 (42.85%) | 24 (28.57%) | 21 (20.79%) | 28 (32.18%) |
| CP-introduced | 33 (71.73%) | 11 (52.38%) | 21 (50%) | 12 (50%) | 12 (57.14%) | 21 (75%) |

- ❖ mostly function words and very common function words
 - ❖ *⟨always,alltid,sempre⟩*

Evaluation: scenario 2

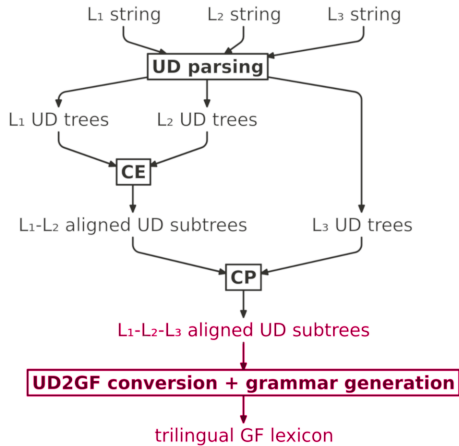
Texts in the same domain (course plans corpora)

| | sv-en-it | it-en-sv |
|---------------|--------------|--------------|
| extracted | 1950 | 1823 |
| propagated | 205 (10.51%) | 200 (10.97%) |
| tot. errors | 66 (32.19%) | 61 (30.5%) |
| CP-introduced | 33 (50%) | 33 (54.09%) |

- ❖ domain-specific concepts
 - ❖ $\langle \text{skills, färdigheter, capacità} \rangle$, $\langle \text{exam, tentamen, prova} \rangle \dots$
 - ❖ $\langle \text{the aim of the course, syftet med kursen, l'obiettivo del corso} \rangle$
 - ❖ an interesting error: $\langle \text{learning, inlärning, conoscere} \rangle$

MT experiments

What's left



From UD to GF alignments

- ❖ UD alignment postprocessing:

- ❖ normalization

```
1  sadly  sadly  ADV  _  _  5  advmod  _  _  
2  this   this   DET  _  _  4  det     _  _  
3  malformed  malformed  ADJ  _  _  4  amod   _  _  
4  subtree  subtree  NOUN  _  5  nsubj  _  _  
5  needs   need   VERB  _  _  0  root   _  _  
6  postprocessing  postprocess  VERB  _  _  5  xcomp
```



```
1  this   this   DET  _  _  3  det     _  _  
2  malformed  malformed  ADJ  _  _  3  amod   _  _  
3  subtree  subtree  NOUN  _  _  0  root   _  _
```

- ❖ selection based on size and usefulness
 - ❖ conversion of UD trees into GF ASTs via `gf-ud`
 - ❖ dependency configurations

From alignments to a grammar

- ❖ aligned ASTs used to automatically generate a GF translation lexicon
- ❖ again via one of gf-ud's modules
 - ❖ requires: **extraction grammar**, **morphological dictionaries**
- ❖ grammar generating simple sentences, limited variation:
 - ❖ *the sentence is simple*
 - ❖ *a sentence is simple*
 - ❖ *sentences are simple*
 - ❖ *these sentences are simple*
 - ❖ *this sentence is an example*
 - ❖ *this short sentence is simple*
 - ❖ *this sentence of the text is simple*

Extending the grammar

Easy to add RGL categories and functions to allow more variation:

- ❑ *this sentence isn't simple*
- ❑ *is this sentence simple?*
- ❑ *this sentence was simple*
- ❑ *this sentence will be simple*
- ❑ *this sentence is simpler than that sentence*

Extending the grammar

Easy to add RGL categories and functions to allow more variation:

- ❑ *this sentence isn't simple*
- ❑ *is this sentence simple?*
- ❑ *this sentence was simple*
- ❑ *this sentence will be simple*
- ❑ *this sentence is simpler than that sentence*

Combining variations:

- ❑ *won't these short sentences be simpler than that long sentence?*

Evaluation: strategy

- ❖ small course plans corpora → 2 bilingual lexica instead of a trilingual one
- ❖ still small lexica + parsing issues → sentences to translate generated in the GF shell
 - ❖ partly arbitrary 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 choices

Evaluation: results

| | DMI (en-it) | CSE (en-sv) |
|-------------|-------------|-------------|
| BLEU-1 to 4 | 55.4 | 61.27 |
| BLEU-1 to 3 | 62.75 | 67.77 |
| BLEU-1 to 2 | 70.6 | 74.3 |
| BLEU-1 | 79.33 | 80.99 |

- ❖ max score:
 - ❖ *⟨the library provides useful textbooks, la biblioteca fornisce libri utili⟩*
 - ❖ *⟨this lab is more difficult than the exam, den här laborationen är svårare än tentamen⟩*
- ❖ min score:
 - ❖ *⟨the test is oral, la prova è dura⟩*
- ❖ most errors are semantical, but 10% of the translation to Italian and 6% of those to Swedish only contain grammatical errors

Conclusions and future work

Conclusions

- ❖ developed a syntax-based CA module
 - ❖ Haskell library + easy to use and configure executables + evaluation scripts
- ❖ evaluation
 - ❖ against a baseline algorithm and a standard statistical tool
 - ❖ in a simple rule-based MT system

Future work