Building Bilingual Multiwords Expressions from Parallel Text

Ramona Enache

Dept. Of Computer Science and Engineering
Chalmers and Univ. of Gothenburg
Based on joint work with Inari Listenmaa and Prasanth Kolachina

Background

- A **multiword expression** (MWE) is a set of at least two words, the meaning of which does not result from composing the meaning of the components

Background

- Important for:
  - machine translation
  - information extraction
  - semantic analysis

Background

- Important for:
  - **machine translation**
  - information extraction
  - semantic analysis

Tool

- Grammatical Framework (GF)

  - type-theoretical grammar formalism

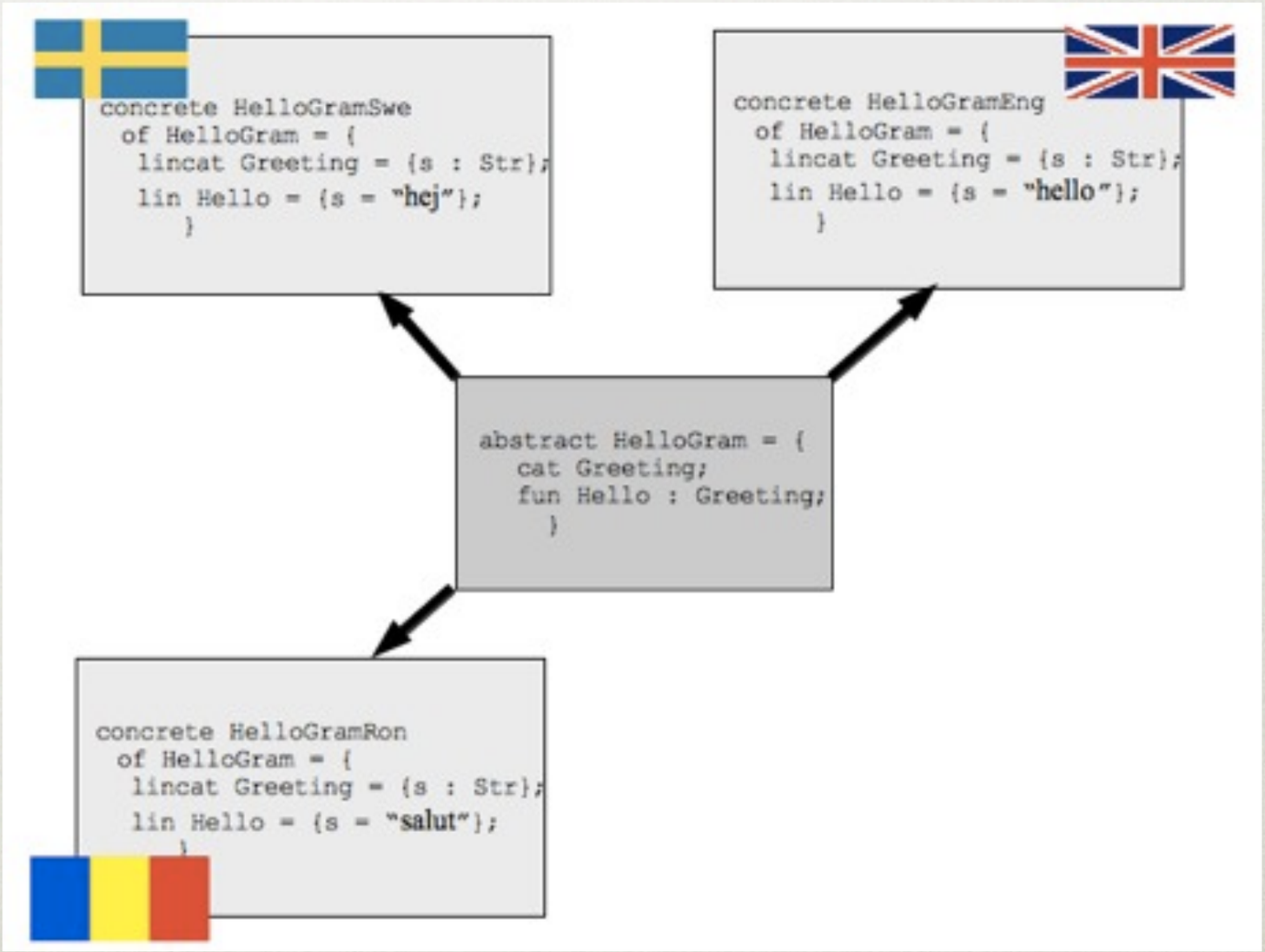
  - dependently-typed functional language for grammar programming

Grammatical Framework

- A GF grammar is
  - an abstract component describing the semantics
  - a number of concrete components mapping the semantics to target languages

Grammatical Framework

Example:



Grammatical Framework

- GF Resource Library

  - abstract grammar - basic syntactic constructions of natural language(predication, complementation, etc) + test lexicon

  - 24 concrete grammars

    - corresponding to languages of the world

    - in addition to syntactical constructions, they feature paradigms for inflectional morphology

Grammatical Framework

GF Resource Grammar:

- computational grammar for the language
- general-purpose
- can be used as a library for domain-specific grammars

Grammatical Framework

GF grammars for natural languages:

- syntactically-correct

- precise w.r.t the semantic from the abstract

- rule-based translation systems between any pair of languages for which a concrete description exists

Grammatical Framework

GF grammars for natural languages:

- syntactically-correct

- precise w.r.t the semantic from the abstract

- rule-based translation systems between any pair of languages for which a concrete description exists

Grammatical Framework

Problems with GF grammars:

- usually built manually

- strict coverage

- compositional translation systems (literal)

Grammatical Framework

Problems with GF grammars:

- **usually built manually**

- strict coverage

- compositional translation systems (literal)

Automated GF Grammars

Building GF grammars in a more automated manner

- from examples

- extracting multilingual single-word lexicons from parallel corpus (phrase tables)

Grammatical Framework

Problems with GF grammars:

- **usually built manually**

- strict coverage

- **compositional translation systems (literal)**

GF and MWEs

GF-style definition for (bilingual) MWEs:

The pair (m1, m2), where:

- at least one of them contains at least two words

- m1 is a (possible) translation of m2

- the translation is not obtained compositionally

=> in most cases, at least one of them is a MWE(according to the first definition)

GF and MWEs

GF-style definition for MWEs:

Example:

- how old are you?/quelle âge as-tu? *(what age do you have?)*

- what is your name?/ comment t-appelles tu? *(how do you call yourself?)*

GF and MWEs

A special case of MWEs emerged -

Compound words

- carboxylic acid/karboxylsyra (Swe)
- blood substitute/blutersatz (Ger)

GF and MWEs

Advantages

- generate all other declension forms and map them

  carboxilic acids/karboxylsyror

- generalize relational MWEs

  how old is your daughter?/quelle âge a ta fille ?

GF and Compounds

- General method
- Use case: English + German
  - Domains: biomedical patents, Europarl
  - Additional constraints:
    - NPs
    - extract from phrase-tables

GF and Compounds

General method

- German rules for compounding, specified as new grammar:

  + w1 + lowercase(w2)
  + w1 + ‘s’ + lowercase(w2)

  ...

GF and Compounds

General method

- extract candidate pairs from phrase tables

+ above a confidence threshold

+ English parses as NP

+ one word for German

GF and Compounds

General method

- greedy algorithm for splitting the German word in the smallest number of constituents that appear in the monolingual dictionary (based on Wiktionary)

GF and Compounds

General method

- use the pair of parse trees to create an entry in a static dictionary

GF and Compounds

Example

abdominal surgery / Bauchchirurgie

+ 2 words English, 1 word German

+ English parses as NP

+ German word not found in dictionary

- first split: Bauchch + irurgie

- ...

- found Bauch + Chirurgie as best match

GF and Compounds

Evaluation

hard to measure precision and recall for large phrase tables due to lack of grammaticality

=> parse with TAG parser, keep NP-only English bits and their German counterparts

GF and Compounds

Evaluation

- 91% of the English phrases parsed in GF

- 73% of the German compounds matched (more rules, proper names)

GF and bilingual MWEs

General method

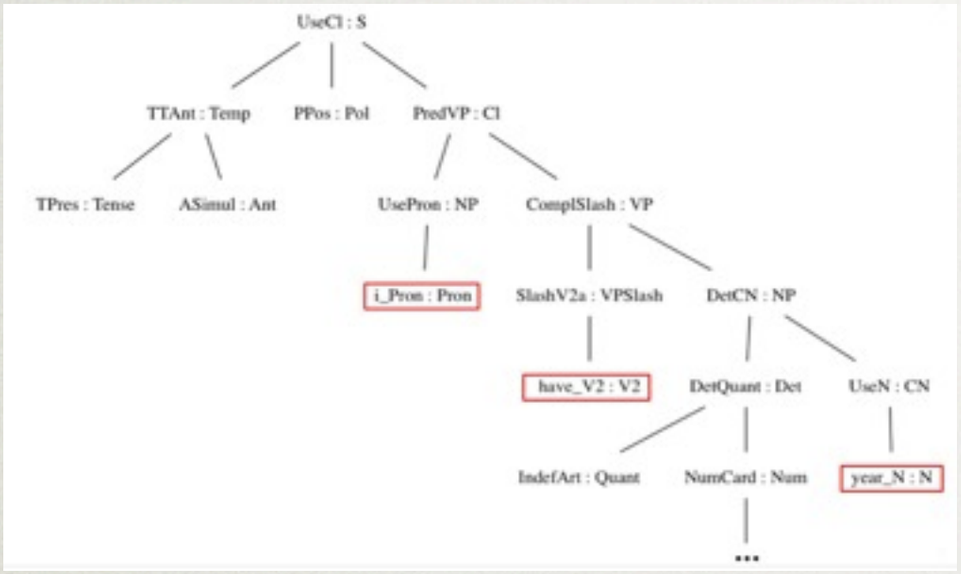
- parse pairs of sentences

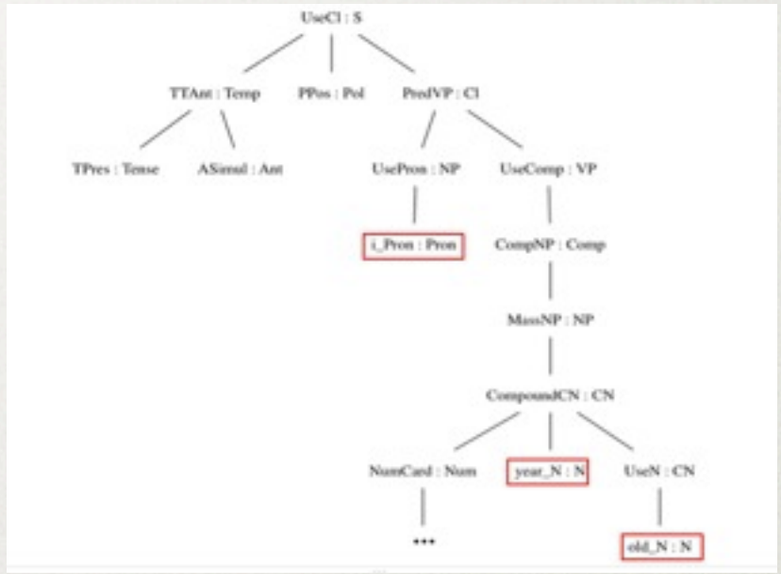
- get trees with least edit distance

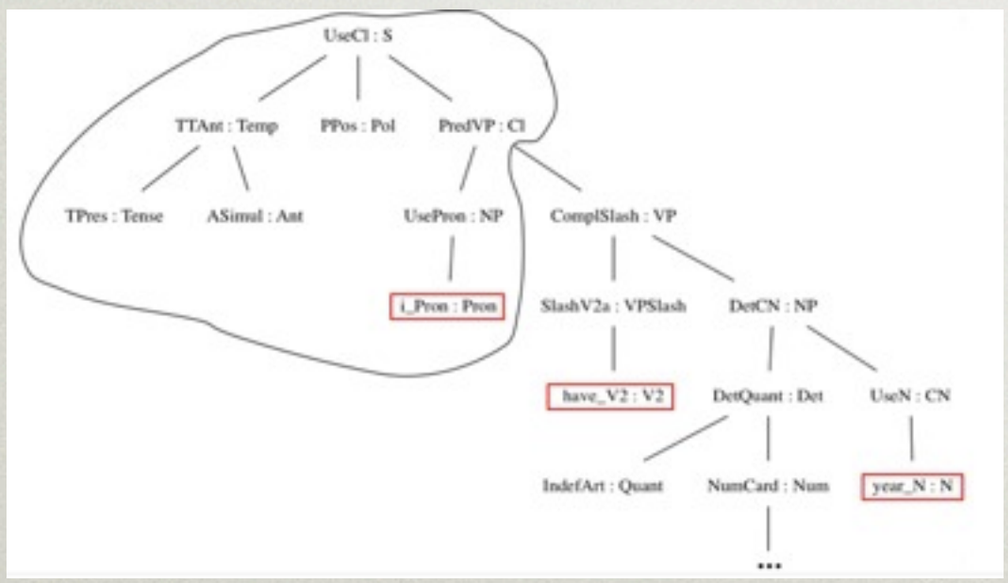
- prune top-down to find differences

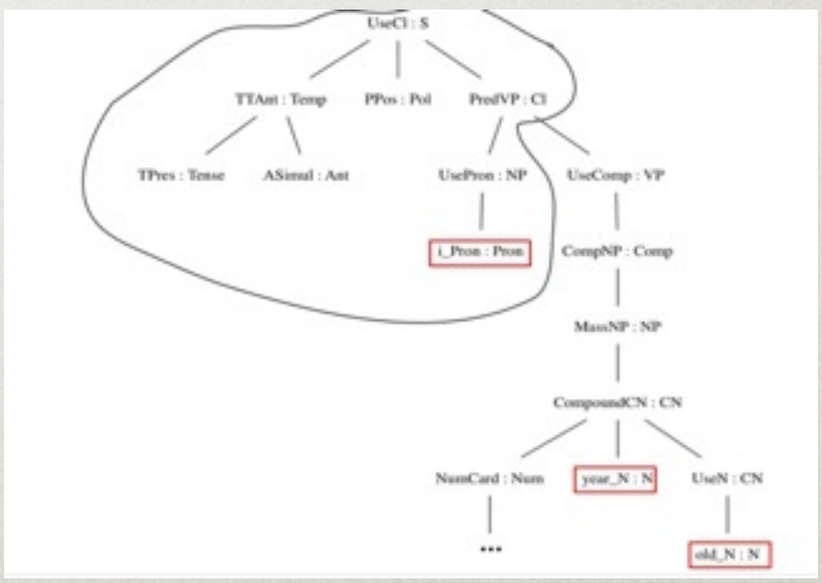
- generalize bottom-up when applicable

GF and bilingual MWEs









GF and bilingual MWEs

Evaluation

- hard to on free text, because of the lack of robustness and quality of existing bilingual GF lexicons

GF and bilingual MWEs

Evaluation

- used to extend grammar semi-automatically from simpler parallel data like **tourist phrases** (Wikitravel)

GF and bilingual MWEs

Future work

- extract MWE from free text, not via phrase tables - can capture discontinuous constituents

- express boundaries for non-compositionality *(kick the bucket/ the bucket was kicked by him)*

- infer compound formation for new words, based on analyzing collected data

- evaluate impact on GF-driven machine translation

Thank you for your attention!