Lexicon extraction in MOLTO

Krasimir Angelov, Lauri Carlson, Ramona Enache, Inari Listenmaa, Aarne Ranta, Shafqat Virk

Overview

1. Introduction
2. Types of lexicons
3. Lexicon sources
4. Showcases
5. Future work

Introduction

Lexicon

- GF lexicon is a part of grammar: we need
  - baseform
  - inflection paradigm
  - valency frame

Use cases

- Converting existing resources to GF lexicons
  - Need to produce mappings from source formats to GF
- Managing terms on TermFactory
  - TermFactory: terminology management platform developed in UHEL
  - Automatic conversion from TermFactory format to GF
  - To be combined with translator's tools

Types of lexicons

Monolingual

Multilingual

- Uni-sense
- Multi-sense

Monolingual lexicons

- Extracted from a monolingual lexicon
- Idiomatic, tailored for each language
- Used as a resource or for a monolingual application

```
DictEng.gf:

    a_priori_Adv = mkAdv "a priori";
    aardvark_N = mkN "aardvark" ;
    ab_initio_Adv = mkAdv "ab initio";
    aback_Adv = mkAdv "aback";
    abactinal_A = mkA "abactinal" ;
    abandon_V2 = mkV2 (mkV "abandon");
```

```
DictGer.gf:

    a_priori_Adv = mkAdv "a priori" ;
    aachener_N = reg2N  "Aachener" "Aachener" masculine ;
    aal_N = reg2N  "Aal" "Aale" masculine ;
    aalfang_N = reg2N  "Aalfang" "Aalfänge" masculine ;
    aasvogel_N = reg2N  "Aasvogel" "Aasvögel" masculine ;
    abaenderbar_A = regA "abänderbar" ;
```

Multilingual lexicons

- Common abstract syntax, concrete syntaxes are translations of it
- Used for multilingual application
- Possible problems
  - Idiomatic POS differences (compound word vs. adjective modifier)
  - Exact word sense matching

```
DictEngGer.gf:

    abandon_V2 = dirV2 (irregV "verlassen" "verlasst"
                               "verließ" "verließe" "verlassen" );
    abase_V2 = dirV2 (irregV "erniedrigen" "erniedrigt"
                             "erniedrigte" "erniedrigte" "erniedrigt");
    abasement_N = mkN "Erniedrigung";
```

Uni-sense lexicons

- Uni-sense: one-to-one correspondence between source and target
- Benefits of uni-sense
  - Lightweight, simple
  - Good results in many cases
    - Carlson and Lindén, 2010: "80 % of the mappings are one-to-one and unproblematic" (Finnish translation of WordNet)
- Problems of uni-sense
  - Arbitrary choice of word sense for the remaining 20 %

Multi-sense lexicons

- Synsets from WordNet, every distinct word sense gets an entry
- Possible to combine with external word sense disambiguation tool
- Example:

```
LinkedDictGer:
    brother_08111676_N = reg2N "Bruder" "Brüder" masculine ;
    brother_08112052_N = reg2N "Bruder" "Brüder" masculine ;
    brother_08112265_N = reg2N "Kamerad" "Kameraden" masculine ;
```

- Format: *lemma*_**senseNumber**_POS
- Sense numbers independent of language

Multi-sense lexicons

Option 1: All synonyms in abstract syntax

- Include every combination of lemma and word sense
  brother_08111676_N
  brother_**08112961**_N
  ...
  buddy_**08112961**_N
- Important for parsing
- Increases the size of the lexicon

Multi-sense lexicons

Option 2: One synonym per word sense

- Include only single word senses, one lemma represents
  brother_08111676_N
  brother_**08112961**_N
  ...
  buddy_**09877951**_N
- Enough for linearization purposes
- Option chosen for LinkedDict.gf

Source formats

Annotated data

- TermFactory RDF
  - Includes necessary information for GF grammars
  - Conversion from TF RDF to GF included in the platform
- WordNet
- Morphological lexicons
  - Bulgarian: OpenOffice spellcheck
  - Finnish: Nykysuomen sanalista by the Institute for the Languages of Finland
  - French: Morphalou
  - Swedish: Folkets lexikon

Source formats

Unannotated data

- Domain ontologies
- Phrase tables
  - Phrase table: Entries of (source chunk, target chunk, probability) for a SMT system
  - Learned from parallel data
  - Experiments for French and German using the phrase tables produced in patent case
- Unannotated word lists
  - Bulgarian: Apertium dictionary
  - Urdu: Waseem Siddiqi's English-Urdu dictionary
  - Various languages: Wiktionary

Combining different sources

- DictEng
  - Lexicon exctracted from Oxford Advanced Learner's Dictionary
  - Valency information extracted from Penn Treebank
  - Manual work with small closed classes: prepositions, irregular verbs
- DictEngBul
  - Valencies from DictEng
  - Inflection paradigms from existing DictBul (whose source is the OpenOffice spellcheck)

Showcases

1. TermFactory: term extraction and conversion to GF
2. WordNet: High-coverage lexicons for robust parsing
3. Phrase tables: domain-specific lexicons for hybrid systems

TermFactory

- TermFactory: Platform for terminology management
- RDF format with GF-specific predicates
- Example for Finnish: language-specific conventions to mark valency, mapped to GF constructors

```
term1:fi-koira-N_-_ont-Dog
      syn:frame "N" ;
      gf:lin "mkN str" ;
      term:hasDesignation exp1:fi-koira-N ;
      term:hasReferent ont0:Dog .

term1:fi-aviomies-N_-_ont-Husband
      syn:frame "jonkun N" ;
      gf:lin "mkN2 (mkN str) (casePrep genitive)" ;
      term:hasDesignation exp1:fi-aviomies-N ;
      term:hasReferent ont0:Husband .

term1:fi-kieltää-V_-_sem-Forbid
      syn:frame "V jotakuta olemasta" ;
      gf:lin "mkV2Vf (mkV str) (casePrep partitive) infElat" ;
      term:hasDesignation exp1:fi-kieltää-V ;
      term:hasReferent sem0:Forbid .
```

TermFactory

- `syn:frame`: user-friendly way to annotate valency
- `gf:lin`: GF constructor
- Mapping from `syn:frame` to `gf:lin`:

```
[] gf:mapping
    [ syn:frame "N" ; gf:lin "mkN str" ] ,
    [ syn:frame "jonkun N" ;
      gf:lin "mkN2 (mkN str) (casePrep genitive)" ],
    [ syn:frame "V jotakuta olemasta" ;
      gf:lin "mkV2Vf (mkV str) (casePrep partitive) infElat" ] .
```

- GF format after conversion:

```
ont_Dog = mkN "koira" ;
ont_Husband = mkN2 (mkN "aviomies") (casePrep genitive) ;
sem_Forbid = mkV2Vf (mkV "kieltää") (casePrep partitive) infElat ;
```

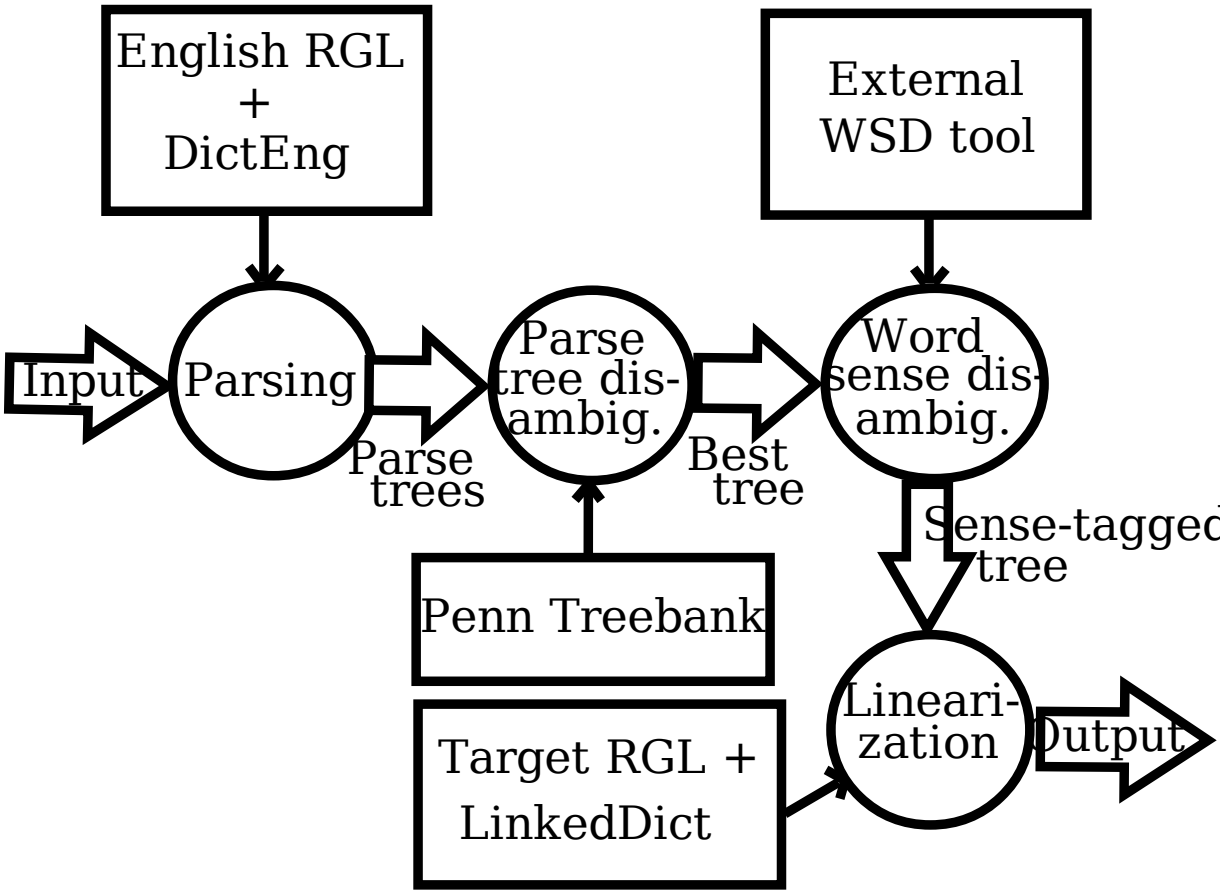
- Demo: to appear in tfs.cc

High-coverage lexicons for robust parsing

Work by Shafqat Mumtaz Virk and K. V. S. Prasad

- Multilingual multi-sense lexicon from WordNets (Hindi and Universal WordNet)
- External word sense disambiguation tool in parsing
- Slight improvement in results for Hindi and German

Parsing and linearization with WSD



Parsing and linearization with WSD

Explanation of the graph:

- Parsing with source language GF resource grammar
- If input is syntactically ambiguous:
  - Multiple parse trees
  - Disambiguation with statistical model built from Penn Treebank data
- Word sense disambiguation of the best tree, with external tool
- Linearization to target language
  - Target language resource grammar
  - Target language LinkedDict

Robust parsing

Work by Krasimir Angelov, Aarne Ranta

- Coverage attained by
  - partial and shallow parsing
  - extended RGL constructions
  - extended lexicon
- Demo and further information in Flagship 3
- Currently only English RGL good enough for parsing
- High-coverage lexicon for Bulgarian, German, Finnish, Hindi, Swedish, Urdu: translations English→*
- Later perhaps translations *↔*
  - Potentially better than SMT that uses English as pivot
- Good for under-resourced languages: SMT needs much more data to produce good results

Phrase tables

Work by Ramona Enache

- Possible to build domain-specific lexicons from unannotated data
- Hybrid systems:
  - get the lexical coverage provided by extensive data
  - maintain the grammaticality provided by RBMT
- Process
  1. English input file
  2. POS-tagging + lemmatization
  3. Lookup in DictEng
  4. Extract valid entries
  5. Obtain translation from (Eng,Ger) phrase-tables
  6. Lookup & Add to DictEngGer

Future work

- Domain-specific tailoring for lexicons
- Better coverage
- More language pairs for robust translation
- Integrating TermFactory to translators' tools

Thank you!