WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– 3rd Project Meeting –

Helsinki, August 31th, 2011

WP5

Overview

1. Overview
2. Ongoing work
3. Future work
4. Dissemination

Overview

Goal

Statistical extension of the grammar-based translation methods to widen their coverage and quality in unconstrained text translation

Overview

Goal

Statistical extension of the grammar-based translation methods to widen their coverage and quality in unconstrained text translation

Especially related to:

WP2 Grammar-based translation method

WP7 Quasi-unconstrained domain, patents

WP9 Evaluation

Overview

Participants & PMs & Tasks

UPC	**38**	SMT technology, hybrid models, corpora processing, evaluation

Overview

Participants & PMs & Tasks

UPC	**38**	SMT technology, hybrid models, corpora processing, evaluation
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT

Overview

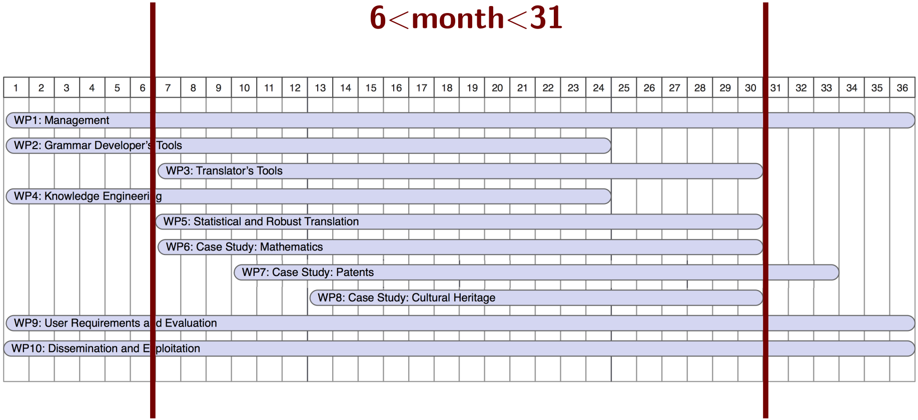
Participants & PMs & Tasks

UPC	**38**	SMT technology, hybrid models, corpora processing, evaluation
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT
UHEL	**6**	Usability and evaluation of the combined system

Overview

Timeline

6<month<31



Overview

Milestones & Deliverables

Month 18 — Month 24 — Month 30

MS5
First prototypes of the *baseline* combination models

D51
Description of the final collection of corpora

Overview

Month 18

Deliverable 5.1

Description of the final collection of corpora

- Work in progress -UPC-: draft version on the web (comments more than welcome!)
- Still, provisional version with parallel corpus extracted and prepared from MAREC corpus
- Preliminary data from WP7

Overview

Month 18

Milestone S5

First prototypes of the *baseline* combination models

- SMT baseline -UPC-: built with current corpus
- GF baseline -GOT-: a first version is available (working day work!)
- Combination baseline -UPC-: to be done

Overview

Month 18

Milestone S5

First prototypes of the *baseline* combination models

But... some hybrid approaches have been explored:

- Combination of GF and SMT alignments
- Lexicon building (translation)

Milestones & Deliverables

Month 18 — **Month 24** — Month 30

MS7
First prototypes of hybrid combination models

D52
Description and evaluation of the combination prototypes

Month 18 — Month 24 — Month 30

MS8
Translation tool complete

D53
WP5 final report: statistical and robust MT

Ongoing work

Overview

1. Overview
2. Ongoing work
   - Scheduled plan
   - Baselines
3. Future work
4. Dissemination

Ongoing work

Tasks

5.1 Parallel corpus compilation in Patents domain

5.2 Out-of-domain corpora

5.3 Synthetic corpora generation **?**

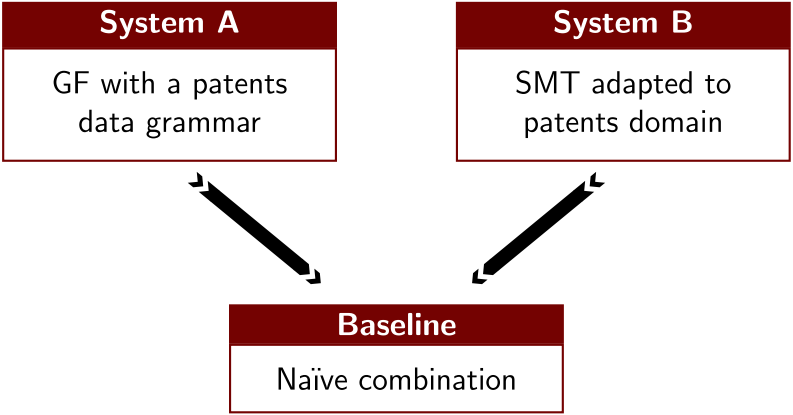
5.4 Baseline systems

5.5 Hybrid Models

5.6 Evaluation of systems

Ongoing work

Baseline systems (Ongoing work: System A)



System A

GF with a patents data grammar

System B

SMT adapted to patents domain

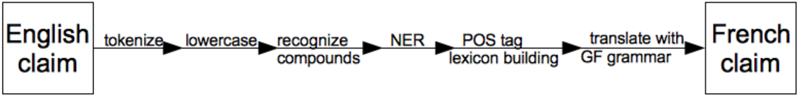
Baseline

Naïve combination

Ongoing work

GF With a patents data grammar

English-to-French patent translator



Ongoing work

GF With a patents data grammar

Pipeline: generic processing

- On-purpose **tokenizer** for treating compound noun phrases separated by hyphens, chemical compounds, etc.
- Stanford POS-tagger for **Named entities** recognition
- **Number** recognizer
- **Chemical compounds** processing

Ongoing work

GF With a patents data grammar

Pipeline: Lexicon Building

- GF library multilingual **lexicon extended** with nouns, adjectives, verbs and adverbs
- **Abstract syntax** for these PoS is created from the claims in English
- **Lemmatisation** and **manual correction** from noise and ambiguities

Ongoing work

GF With a patents data grammar

Pipeline: Lexicon Building II

- **Inflection** generated using the implemented GF paradigms and the English dictionary of the GF library
- **Base forms** are **translated** into French and the inflection is generated in the same way

(Future extension to other languages)

Ongoing work

GF With a patents data grammar

Pipeline: Grammar

- Extension of the Resource Grammar with functions implementing constructions that occur in patent claims
- Huge number of ambiguities
- For the moment, the coverage is around 15% on complete sentences

Future work

Overview

1 Overview

2 Ongoing work

3 Future work

4 Dissemination

Future work

Related to GF baseline

Increase parser **robustness** by

- Chunking the claims and parsing the chunks separately
- Recombine the results with the help of the grammar

Reduce **ambiguity** by

- bottom up disambiguation based on the corpus

Future work

Related to GF baseline

Widen grammar **coverage** by

- Write more rules
- Detect idioms (latin expressions, law jargon)
- Detect prepositions and conjunctions which are specific to patents and extend the lexicon with them

Future work

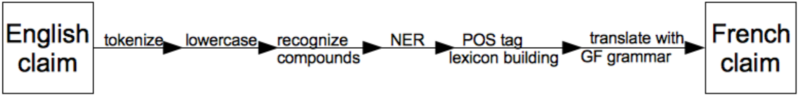
Related to SMT baseline

- Build a new **corpus** (if we're lucky!)
- Train the SMT system and obtain **translation models** with the new corpus
- Automatic **evaluation** and comparison with the GF baseline

Future work

Related to GF+SMT

Evaluation and **homogenisation** of the GF and the SMT baseline pipeline



English claim

tokenize

lowercase

recognize compounds

NER

POS tag lexicon building

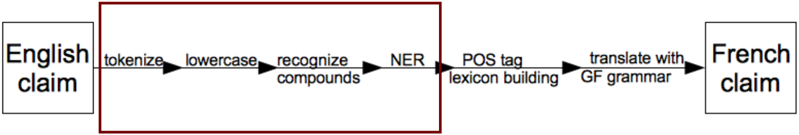
translate with GF grammar

French claim

Future work

Related to GF+SMT

Evaluation and **homogenisation** of the GF and the SMT baseline pipeline



GF+SMT

Future work

Related to GF+SMT

- **Combination baseline**
  Cascade translation sentences and/or chunks
- **Hard integration** GF+SMT
  Force fixed GF translations within a SMT system
- A first automatic **evaluation** of the resulting systems

Dissemination

Overview

1. Overview
2. Ongoing work
3. Future work
4. Dissemination

WP dissemination

MOLTO Papers

- **Patent translation within the MOLTO project**
  Cristina España-Bonet, Ramona Enache, Adam Slaski, Aarne Ranta, Lluís Màrquez and Meritxell Gonzàlez
  MT Summit XIII 4th Workshop on Patent Translation. Xiamen, September 2011

MOLTO Report

- **Towards a RB-SMT Hybrid System for Translating Patent Claims – Results and Perspectives**
  Ramona Enache and Adam Slaski
  Internal Report.

Dissemination

WP dissemination

Related Papers

- **Hybrid Machine Translation Guided by a Rule-Based System**
  Cristina España-Bonet , Gorka Labaka, Lluís Màrquez, Arantza Díaz de Ilarraza and Kepa Serasola
  MT Summit XIII. Xiamen, September 2011

Anything planned?

WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– 3rd Project Meeting –

Helsinki, August 31th, 2011