WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– Second year review –

Barcelona, March 20th, 2012

WP5

Overview

1. General view
2. Ongoing work
3. Future work
4. Dissemination

General view

Goal

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

General view

Goal

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

General view

Participants & PMs & Tasks

UPC	38

SMT technology, hybrid models, corpora processing, evaluation

General view

Participants & PMs & Tasks

UPC	**38**

SMT technology, hybrid models, corpora processing, evaluation

UGOT	**9**

Extension of GF to open domain, robust parsing

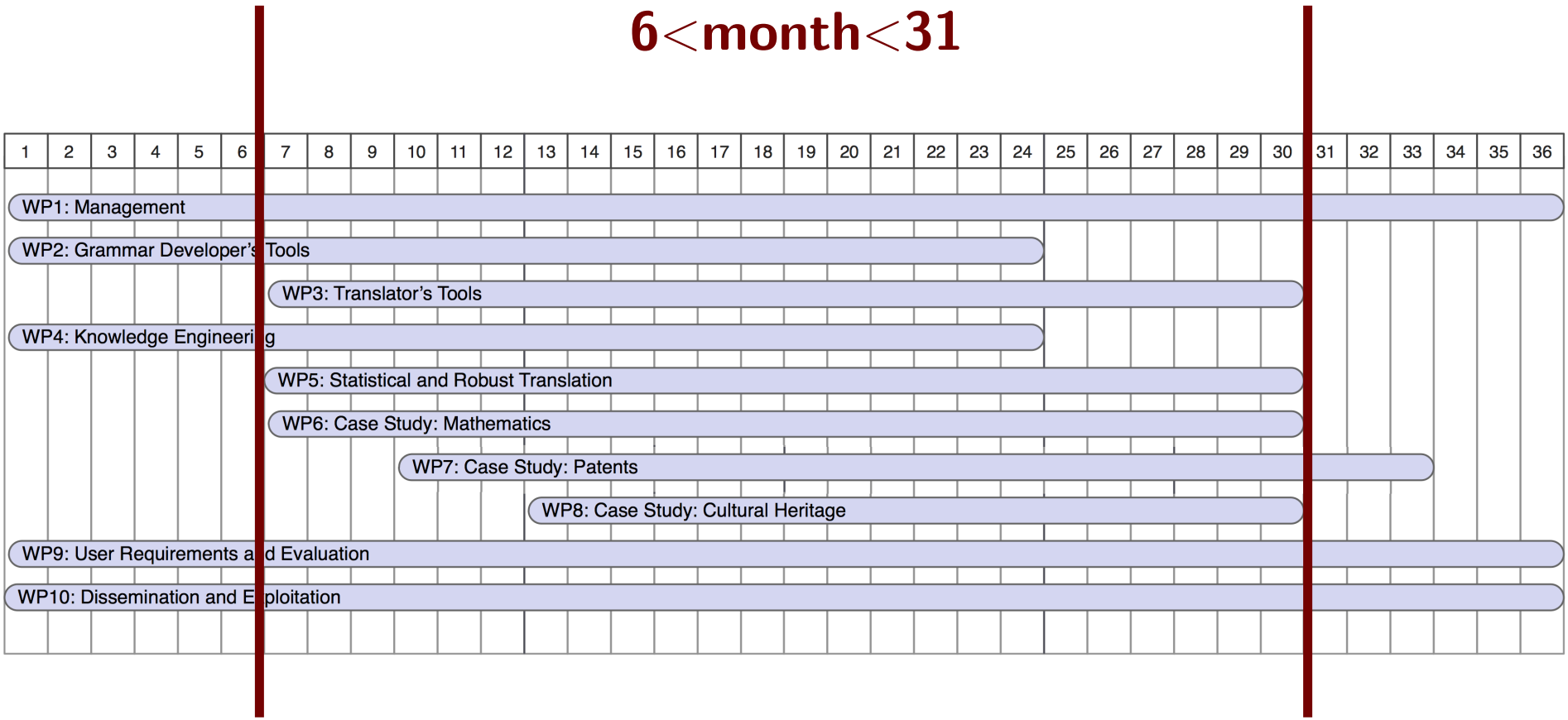
General view

Participants & PMs & Tasks

UPC	**38**	SMT technology, hybrid models, corpora processing, evaluation
UGOT	**9**	Extension of GF to open domain, robust parsing
UHEL	**6**	Usability and evaluation of the combined system

General view

Timeline



General view

Milestones & Deliverables

Month 18 — Month 24 — Month 30

MS5

First prototypes of the *baseline* combination models.

D51

Description of the final collection of corpora.

General view

Milestones & Deliverables

Month 18 — **Month 24** — Month 30

MS7

First prototypes of hybrid combination models.

D52

Description and evaluation of the combination prototypes.

General view

Milestones & Deliverables

Month 18 — Month 24 — **Month 30**

MS8

Translation tool complete.

D53

WP5 final report: statistical and robust MT.

Ongoing work

Overview

Ongoing work

Tasks

T5.1 Parallel corpus compilation in Patents domain

T5.2 Out-of-domain corpus

T5.3 Robust parsing

T5.4 Baseline systems

T5.5 Hybrid Models

T5.6 Systems evaluation

Ongoing work

Tasks status

T5.1 Parallel corpus compilation in Patents domain

D51–M18 Description of the final collection of corpora
But! New corpus at M22

Ongoing work

Tasks status

T5.1 Parallel corpus compilation in Patents domain

D51–M18 Description of the final collection of corpora
But! New corpus at M22

From MAREC to EPO (claims A61P)

SET	Seg DE-EN	Seg FR-EN	Seg FR-DE
Training	279,282	279,282	279,282
Development	993	993	993
Test	1,008	1,008	1,008
NEW EPO	847	858	831

Ongoing work

Tasks status

T5.1 Parallel corpus compilation in Patents domain

From MAREC to EPO (total of claims)

	Seg DE-EN	**Seg FR-EN**	**Seg FR-DE**
NEW EPO	1,757,860	1,750,464	1,695,505

Ongoing work

Tasks status

T5.1 Parallel corpus compilation in Patents domain

From MAREC to EPO (total of claims)

	Seg DE-EN	**Seg FR-EN**	**Seg FR-DE**
NEW EPO	1,757,860	1,750,464	1,695,505

T5.2 Out-of-domain corpus

Subset of the data sets included in WMT11

Ongoing work

Tasks status

T5.3 Robust parsing

General chunk parsing (on PennTreebank).
Core work during 1st year

Chunking adapted to patents.
Included now in GF patent translator

Ongoing work

Tasks status

T5.3 Robust parsing

Chunking for patents

- Considered **chunks**: VP, NP, AP and PP
- VP, RP and AP need to have an NP to **agree** with
- Chunking **adapted** to patents estruture
- Merging

Ongoing work

Tasks status

T5.4 Baseline systems

SMT baseline. Core work during 1st year

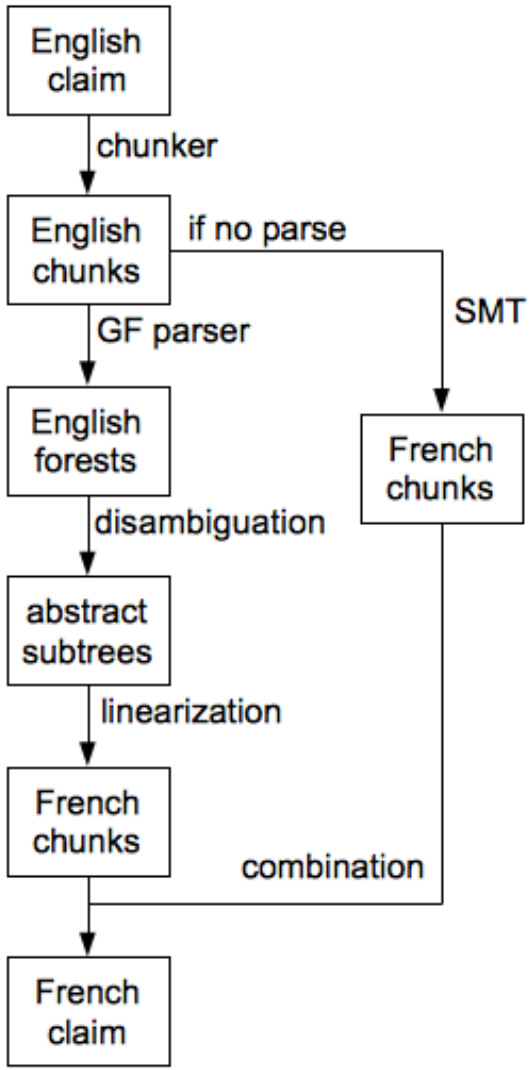
GF baseline. News: Complete GF1 translator

Ongoing work

Tasks status

T5.4 Baseline systems, GF

GF
English-to-French
patent translator



Ongoing work

Tasks status

T5.4 Baseline systems, GF

On-line lexicon building

- Pre-process: English claims tagged **PoS** (Genia)
- Lemmatised with GF **English lexicon**
- New lexicon included as **abstract syntax** entries
- SMT English-to-French **translated lexicon**

Ongoing work

Tasks status

T5.4 Baseline systems, GF

Grammar

- Extension of the Resource Grammar with functions implementing **constructions** that occur in patent claims
- Huge number of ambiguities
  Disambiguation: frequency counts in the corpus
- The **coverage** is of 7% on complete sentences and 33% on chunks

Ongoing work

Tasks status

T5.5 Hybrid Models

1. **Hard** integration

Force fixed GF translations within a SMT system.

2. **Soft** integration led by **SMT**

Make available GF translations to a SMT system.

3. **Soft** integration led by **GF**

Complement with SMT options the GF translation structure.

Ongoing work

Tasks status

T5.5 Hybrid Models, 3. SI

3. GF is complemented with SMT translations

Complement the GF translation structure with SMT options.

The GF baseline system is already hybrid with SMT.

Ongoing work

Tasks status

T5.5 Hybrid Models, 1. HI

1. SMT includes reliable GF translations

Force fixed GF translations within a SMT system.

Characteristics:

- SMT translates untranslated chunks by GF
- Reordering of chunks allowed
- No interaction among chunks (phrases)

Ongoing work

Tasks status

T5.5 Hybrid Models, 2. SI

2. SMT adds GF translations

GF translations are phrases with a probability within SMT system.

Characteristics:

- SMT translates translated & untranslated chunks by GF
- Reordering of chunks allowed
- Interaction among phrases GF and SMT phrases

Ongoing work

Tasks status

T5.6 Systems evaluation

English-to-French hybrid translation

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF	60.96	50.08	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	27.03	17.50	25.32	63.18	9.99	44.58	71.64	72.65	67.14

Ongoing work

Tasks status

T5.6 Systems evaluation

English-to-French hybrid translation

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF	60.96	50.08	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	27.03	17.50	25.32	63.18	9.99	44.58	71.64	72.65	67.14
HI	33.56	21.95	31.24	55.88	9.24	38.81	67.30	67.80	58.84
SI1.0	26.76	17.39	25.10	63.56	10.02	**44.86**	**71.96**	72.89	67.56
SI0.5	**26.63**	**17.32**	**25.02**	**63.60**	**10.03**	44.84	71.94	**72.93**	**67.60**
SI0.0	27.08	17.48	25.36	63.15	9.99	44.54	71.60	72.66	67.11

T5.6 Systems evaluation

Example

GF	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **at the same time as**...
SMT	Utilisation selon la revendication 3, dans laquelle **le médicament séparée** est administré **en même temps que**...

T5.6 Systems evaluation

Example

GF	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **at the same time as**...
SMT	Utilisation selon la revendication 3, dans laquelle **le médicament séparée** est administré **en même temps que**...
HI	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **en même temps que**...
SI0.5	Utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **en même temps que**...
Ref.	Utilisation selon la revendication 3, dans laquelle le médicament **séparé** est administré **en même temps que**...

Ongoing work

Tasks status

T5.6 Systems evaluation

Preliminar manual evaluation (100 fragments in French)

	SMT	Tied	SI0.5
Tester1	4	9	10
Tester2	3	13	7
Tester3	2	17	4
Tester4	6	5	12
Total	15	44	33

Future work

Overview

Future work

Tasks planning

T5.1 Parallel corpus compilation in Patents domain

Complete

T5.2 Out-of-domain corpus

Complete

Future work

Tasks planning

T5.1 Parallel corpus compilation in Patents domain

Complete

T5.2 Out-of-domain corpus

Complete

T5.3 Robust parsing

Include general robust parsing into hybrid systems

Future work

Tasks planning

T5.4 Baseline systems

- SMT and GF on new test set
- SMT out-of-domain translation
- German patents grammar
- GF system 2 (with SMT components before parsing)
- Compounds grammar

Future work

Tasks planning

T5.5 Hybrid systems

- New integration approach (weighting probabilities)
- Integrate GF system 2 with SMT
- Hybrid with robust parsing

Future work

Tasks planning

T5.5 Hybrid systems

- New integration approach (weighting probabilities)
- Integrate GF system 2 with SMT
- Hybrid with robust parsing

T5.6 Systems evaluation

- Final automatic evaluation
- Manual evaluation on a selected test set

Dissemination

Overview

Dissemination

WP dissemination

Refereed Conferences

- *Patent translation within the MOLTO project*
  Cristina España-Bonet, Ramona Enache, Adam Slaski, Aarne Ranta, Lluís Màrquez and Meritxell Gonzàlez
  Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII, Xiamen, China, September 23, 2011.

- *A Hybrid System for Patent Translation*
  Ramona Enache, Cristina España-Bonet, Aarne Ranta and Lluís Màrquez
  Submitted to EAMT 2012.

Dissemination

WP dissemination

Talks

- *GF patent translation system*
  Ramona Enache, Adam Slaski
  GF Summer School, Barcelona, August 2011.

- *SMT within MOLTO's hybrid translation system*
  Cristina España-Bonet
  GF Summer School, Barcelona, August 2011.

- *Introduction to SMT and its standard tools*
  Cristina España-Bonet
  GF Summer School, Barcelona, August 2011.

Dissemination

WP dissemination

Report

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

Related theses

- *The Mechanics of the Grammatical Framework*
  Krasimir Angelov
  Ph.D. Thesis, University of Göteborg, 2011.

- *Automating the development of multilingual grammars*
  Ramona Enache
  Licenciate Thesis, University of Göteborg, 2012.

Dissemination

WP dissemination

Related refereed publications

- *Deep evaluation of hybrid architectures: simple metrics correlated with human judgments*
  Gorka Labaka, Arantza Díaz de Ilarraza, Cristina España-Bonet, Lluís Màrquez and Kepa Serasola
  Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT11).

- *Hybrid Machine Translation Guided by a Rule-Based System*
  Cristina España-Bonet, Gorka Labaka, Arantza Díaz de Ilarraza, Lluís Màrquez and Kepa Serasola
  Proceedings of the 13th Machine Translation Summit, Xiamen, China, September 19-23, 2011.

WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– Second year review –

Barcelona, March 20th, 2012

Extra slides

Tasks status

T5.3 Robust parsing (Chunking for patents)

Word	**PoS Genia**	**Chunk Genia**	**PoS Final**	**Chunk Final**
The	DT	B-NP	DT	B-NP
use	NN	I-NP	NN	I-NP
of	IN	B-PP	IN	I-NP
claim	NN	B-NP	NN	I-NP
1	CD	I-NP	CD	I-NP
,	,	O	,	O
wherein	IN	B-PP	RP	B-RP
said	V	B-VP	DT	B-NP
use	NN	B-NP	NN	I-NP
is	VBZ	B-VP	VBZ	B-VP
intramuscular	JJ	B-ADJP	JJ	I-VP
.	.	O	.	O

Extra slides

Tasks status

T5.4 Baseline systems, GF

Why the limited coverage?

- Punctuation and similar are not considered (31.3%)
- 18.3% of the chunks cannot be **parsed** by the grammar
- 15.5% of the chunks cannot be translated due to **uncomplete lexicon**
- 1.1% cannot be translated because of the missing information about **agreement**

Extra slides

Tasks status

T5.5 Hybrid Models

1. **Hard** integration

Force fixed GF translations within a SMT system.

2. **Soft** integration led by **SMT**

Make available GF translations to a SMT system.

3. **Soft** integration led by **GF**

Complement with SMT options the GF translation structure.

Extra slides

Tasks status

T5.5 Hybrid Models, 3. SI

3. GF is complemented with SMT translations

Complement the GF translation structure with SMT options.

The GF baseline system is already hybrid with SMT.

Extra slides

Tasks status

T5.5 Hybrid Models, 1. HI

1. SMT includes reliable GF translations

Force fixed GF translations within a SMT system.

Characteristics:

- SMT translates untranslated chunks by GF
- Reordering of chunks allowed
- No interaction among chunks (phrases)

T5.5 Hybrid Models, 1. HI

Proportion of chunks

	GF	SMT
NP	2,366 (14.9%)	2,199 (13.8%)
VP	275 (1.7%)	1,302 (8.2%)
AP	1,960 (12.3%)	1,935 (12.2%)
RP	648 (4.1%)	86 (0.5%)
Other	–	5,099 (32.0%)
Total	*5,301 (33.3%)*	*10,621 (66.7%)*

T5.5 Hybrid Models, 1. HI

Example

A use according to claim 3 , wherein the separate medicament is administered at the same time as the said medicament .

Extra slides

Tasks status

T5.5 Hybrid Models, 1. HI

Example

A use according to claim 3 , wherein the separate medicament is administered at the same time as the said medicament .

A use according to claim 3 , wherein the separate medicament is administered <adv> **at the same time** </adv> <adv> **as the** </adv> **said medicament .**

T5.5 Hybrid Models, 1. HI

Example

A use according to claim 3 , wherein the separate medicament is administered at the same time as the said medicament .

<np GF = "Une utilisation"> **A use** </np> <adv GF = "selon la revendication 3"> **according to claim 3** </adv> **,** <rp GF = "dans laquelle"> **wherein** </rp> <np GF = "le médicament séparé"> **the separate medicament** </np> <vp GF = "est administré"> **is administered** </vp> <adv> **at the same time** </adv> <adv> **as the** </adv> <np GF = "ledit médicament"> **said medicament** </np> **.**

Extra slides

Tasks status

T5.5 Hybrid Models, 2. SI

2. SMT adds GF translations

GF translations are phrases with a probability within SMT system.

Characteristics:

- SMT translates translated & untranslated chunks by GF
- Reordering of chunks allowed
- Interaction among phrases GF and SMT phrases

Extra slides

Tasks status

T5.5 Hybrid Models, 2. SI

Example

A use according to claim 3 , wherein the separate medicament is administered at the same time as the said medicament .

Extra slides

Tasks status

T5.5 Hybrid Models, 2. SI

```
separate ||| séparer ||| 0.178571 0.13172 0.0609756 0.0621039 2.718
separate ||| séparé , ||| 0.357143 0.215329 0.00677507 0.011837 2.718
separate ||| séparé ||| 0.241667 0.215329 0.0785908 0.0747782 2.718
separate ||| séparée , ||| 0.206897 0.723653 0.00813008 0.0619939 2.718
...
the separate ||| séparée ||| 0.00446429 0.269526 1 0.391635 2.718
...
medicament is administered ||| médicament est administré ||| 0.7427 ...
medicament is administered ||| médicament est administrée en ||| ...
medicament is administered ||| médicament est administrée ||| 1 0.6110
...
```

the separate medicament ||| **le médicament séparé** ||| GF probability

Extra slides

Tasks status

T5.6 Systems evaluation

Baseline

Deep evaluation al lexical level for the three language pairs (English-German, English-French, German-French)

Extra slides

Tasks status

T5.6 Systems evaluation

English-to-French hybrid translation

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF	60.96	50.08	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	27.03	17.50	25.32	63.18	9.99	44.58	71.64	72.65	67.14

Extra slides

Tasks status

T5.6 Systems evaluation

English-to-French hybrid translation

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF	60.96	50.08	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	27.03	17.50	25.32	63.18	9.99	44.58	71.64	72.65	67.14
HI	33.56	21.95	31.24	55.88	9.24	38.81	67.30	67.80	58.84
SI1.0	26.76	17.39	25.10	63.56	10.02	**44.86**	**71.96**	72.89	67.56
SI0.5	**26.63**	**17.32**	**25.02**	**63.60**	**10.03**	44.84	71.94	**72.93**	**67.60**
SI0.0	27.08	17.48	25.36	63.15	9.99	44.54	71.60	72.66	67.11

T5.6 Systems evaluation

Example

GF	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **at the same time as**...
SMT	Utilisation selon la revendication 3, dans laquelle **le médicament séparée** est administré **en même temps que**...

T5.6 Systems evaluation

Example

GF	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **at the same time as**...
SMT	Utilisation selon la revendication 3, dans laquelle **le médicament séparée** est administré **en même temps que**...
HI	**Une** utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **en même temps que**...
SI0.5	Utilisation selon la revendication 3, dans laquelle **le médicament séparé** est administré **en même temps que**...
Ref.	Utilisation selon la revendication 3, dans laquelle le médicament **séparé** est administré **en même temps que**...