WP5
Statistical and Robust Translation

Cristina España-Bonet
Lluís Màrquez
TALP Research Center

1st MOLTO Project Meeting

Varna, September 9th, 2010

WP5

Overview

1. General view
2. Planning for WP's first year
3. Hybrid approaches
4. Short term tasks

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:

WP3 Grammar-based translation method.

WP7 Quasi-unconstrained domain, patents.

WP9 Evaluation.

General view

Participants & PMs & Tasks

UPC	**32**	SMT technology, hybrid models, corpora processing.

General view

Participants & PMs & Tasks

UPC	**32**	SMT technology, hybrid models, corpora processing.
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT.

General view

Participants & PMs & Tasks

UPC	**32**	SMT technology, hybrid models, corpora processing.
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT.
?	**6**	Corpora provider.

General view

Participants & PMs & Tasks

UPC	**32**	SMT technology, hybrid models, corpora processing.
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT.
?	**6**	Corpora provider.
UHEL	**3**	Usability and evaluation of the combined system.

General view

Work plan & Participants

UPC	**32**
UGOT	**9**
?	**6**
UHEL	**3**

1. Probabilistic extension of a GF domain grammar.

2. Adapt base SMT systems to the Patents domain.

3. Develop and test hybrid GF-SMT translation methods.

General view

Work plan & Participants

UPC	32
UGOT	**9**
?	6
UHEL	3

1. **Probabilistic extension of a GF domain grammar.**

2. Adapt base SMT systems to the Patents domain.

3. Develop and test hybrid GF-SMT translation methods.

General view

Work plan & Participants

UPC	**32**
UGOT	9
?	**6**
UHEL	3

1. Probabilistic extension of a GF domain grammar.

2. **Adapt base SMT systems to the Patents domain.**

3. Develop and test hybrid GF-SMT translation methods.

General view

Work plan & Participants

UPC	**32**
UGOT	**9**
?	**6**
UHEL	**3**

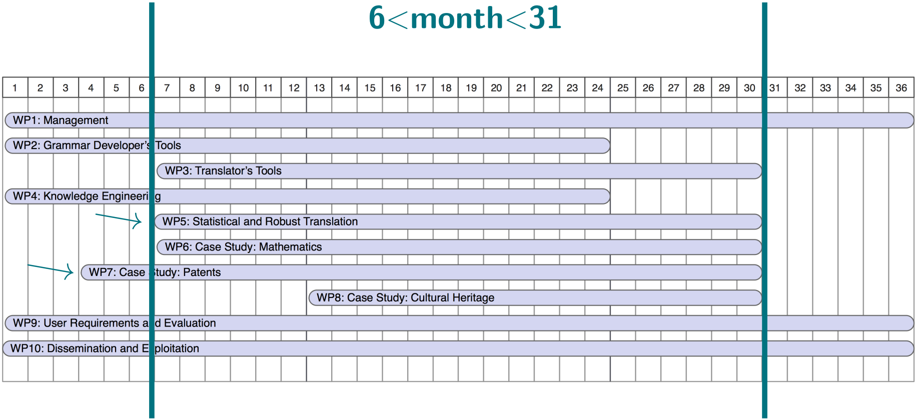
1. Probabilistic extension of a GF domain grammar.

2. Adapt base SMT systems to the Patents domain.

3. Develop and test hybrid GF-SMT translation methods.

General view

Timeline



6<month<31

WP1: Management

WP2: Grammar Developer's Tools

WP3: Translator's Tools

WP4: Knowledge Engineering

WP5: Statistical and Robust Translation

WP6: Case Study: Mathematics

WP7: Case Study: Patents

WP8: Case Study: Cultural Heritage

WP9: User Requirements and Evaluation

WP10: Dissemination and Exploitation

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.

Planning for WP's first year

From month 7 to month 18

First proposal

- Compilation and annotation of corpora from the patents domain.
- Training and adaptation of the base SMT systems.
- Statistical extension of the patents GF grammar.
- Evaluation and comparison of GF, SMT and cascade systems (baselines) in real domain data.
- First experiments with the combination approaches.

Planning for WP's first year

From month 7 to month 18

First proposal

BUT!

- Compilation and annotation of corpora from the **patents domain**.
- Training and **adaptation** of the base SMT systems.
- Statistical extension of the **patents GF grammar**.
- Evaluation and comparison of GF, SMT and cascade systems (baselines) in real **domain data**.
- First experiments with the combination approaches.

Planning for WP's first year

Patents data perspectives

A temporal solution

IRF membership has allowed access to **CLEF-IP 2010** data:

- Test set containing EPO patents.
- Languages: English, French and German.

Planning for WP's first year

Patents data perspectives

A temporal solution

IRF membership has allowed access to CLEF-IP 2010 data:

- Test set containing EPO patents.
- Languages: English, **French** and German.

Minor drawbacks:

- Too small corpus (to be confirmed).
- Languages: English, **Spanish** and German.

Planning for WP's first year

Patents data consequences

In terms of time

WP7 (Case study: Patents) start: Month 4

WP5 (Statistical and Robust translation) start: Month 7

But, first data: Month 8 (at best!)

Planning for WP's first year

Patents data consequences

In terms of time

WP7 (Case study: Patents) start: Month 4

WP5 (Statistical and Robust translation) start: Month 7

But, first data: Month 8 (at best!)

! 4 months minimum delay.

Planning for WP's first year

Patents data consequences

In terms of tasks

An obvious delay in corpora compilation and annotation.

Change of approach:

from optimising base systems to dig into the hybrid system.

Planning for WP's first year

Patents data consequences

In terms of tasks

An obvious delay in corpora compilation and annotation.

Change of approach:

from optimising base systems to dig into the hybrid system.

! Mainly, just a change of order in tasks.

Planning for WP's first year

Patents data consequences

In terms of milestones & deliverables

MS5 First prototypes of the *baseline* combination models.

D51 Description of the final collection of corpora.

Sept. 2011. We can be optimistic if CLEF-IP data is representative and we get the full corpus... before the end of the year?

Planning for WP's first year

Patents data consequences

In terms of milestones & deliverables

MS5 First prototypes of the *baseline* combination models.

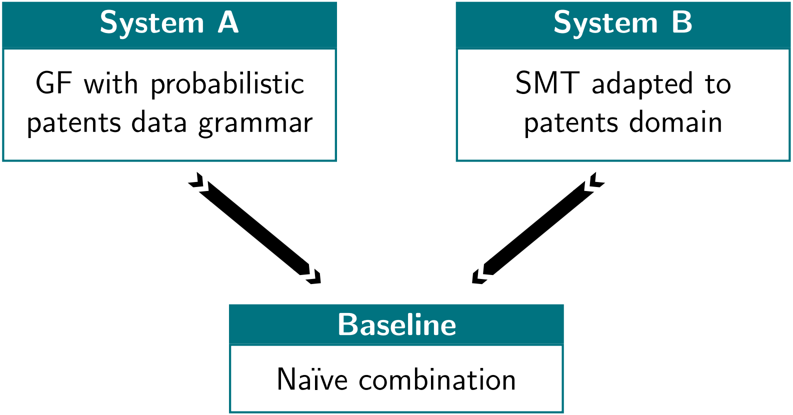
D51 Description of the final collection of corpora.

Sept. 2011. We can be optimistic if CLEF-IP data is representative and we get the full corpus... before the end of the year?

! We are OK.

Hybrid approaches

Step 1: Base and baseline systems



System A

GF with probabilistic patents data grammar

System B

SMT adapted to patents domain

Baseline

Naïve combination

Hybrid approaches

Step 2: Real hybridisation

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.

Hybrid approaches

Hard integration

Force fixed GF translations within a SMT system.

✓ Straightforward to implement from the SMT pov.

◊ Need of GF partial translations.

◊ Waiting for domain adapted base systems.

✗ There is no interaction between GF and SMT.

Hybrid approaches

Soft integration led by SMT (I)

Make available GF translations to a SMT system. (I)

Translation Table, core of an SMT system:

```
source language ||| target language ||| probabilities

...
quite a burden ||| un estorbo muy grande ||| 0.25 1.57587e-06 0.25 3.57895e-12 2.718
quite a burden ||| un estorbo muy ||| 0.25 1.57587e-06 0.25 8.38161e-08 2.718
quite a challenge but we ||| todo un reto , pero lo ||| 0.5 6.64558e-05 1 1.46764e-06 2.718
quite a challenge but ||| todo un reto , pero ||| 0.5 0.00179307 1 9.70607e-05 2.718
quite a challenge ||| todo un reto , ||| 0.5 0.002396 0.5 0.000190619 2.718
quite a challenge ||| todo un reto ||| 0.333333 0.002396 0.5 0.00244338 2.718
quite a considerable delay ||| un retraso muy considerable ||| 0.333333 2.91692e-05 ...
quite a contribution towards ||| una importante contribución en lo ||| 0.25 9.69758e-07 ...
quite a contribution towards ||| una importante contribución en ||| 0.142857 9.69758e-07 ...
quite a difference whether ||| muy diferente ||| 0.0344828 8.29695e-09 1 0.0013126 2.718
quite a difference ||| muy diferente ||| 0.0344828 1.38144e-05 1 0.0013126 2.718
...
```

Hybrid approaches

Soft integration led by SMT (I)

GF scored partial output as **new features** in SMT decoding.

$$\log P(e|f) \sim \lambda_{lm} \log P(e) + \lambda_{g} \log P(f|e) + \lambda_{d} \log P(e|f)$$
$$+\lambda_{di} \log P_{di}(e,f) + \lambda_{w} \log w(e) + \lambda_{\mathbf{GF}} \mathbf{\log P_{GF}(e|f)}$$

`quite a challenge|||todo un reto|||0.333 0.002 0.5 0.002 2.718` $\mathbf{\log P}_{\mathrm{GF}}(e|f)$

Hybrid approaches

Soft integration led by SMT (I)

GF scored partial output as **new features** in SMT decoding.

$$\log P(e|f) \sim \lambda_{lm} \log P(e) + \lambda_g \log P(f|e) + \lambda_d \log P(e|f)$$
$$+\lambda_{di} \log P_{di}(e,f) + \lambda_w \log w(e) + \lambda_{\mathbf{GF}} \mathbf{log\ P_{GF}(e|f)}$$

`quite a challenge|||todo un reto|||0.333 0.002 0.5 0.002 2.718` $\mathbf{log\ P}_{\mathrm{GF}}(e|f)$

Requirements:

- GF predictions have to be probabilistic.
- Phrase pairs without prediction must be complemented.

Hybrid approaches

Soft integration led by SMT (II)

Make available GF translations to a SMT system. (II)

GF and SMT translation options drawn from different sources.

Hybrid approaches

Soft integration led by SMT (II)

Make available GF translations to a SMT system. (II)

GF and SMT translation options drawn from different sources.

⌄

The intersection is only a subgroup of phrases.

Hybrid approaches

Soft integration led by SMT (II)

Make available GF translations to a SMT system. (II)

GF and SMT translation options drawn from different sources.

The intersection is only a subgroup of phrases.

Define three translation tables.

Hybrid approaches

Soft integration led by SMT (II)

GF generated corpus

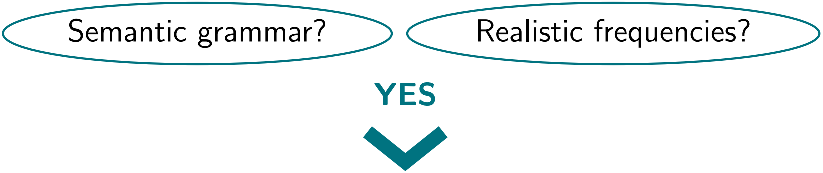
Semantic grammar?

Realistic frequencies?

Hybrid approaches

Soft integration led by SMT (II)

GF generated corpus



Semantic grammar?

Realistic frequencies?

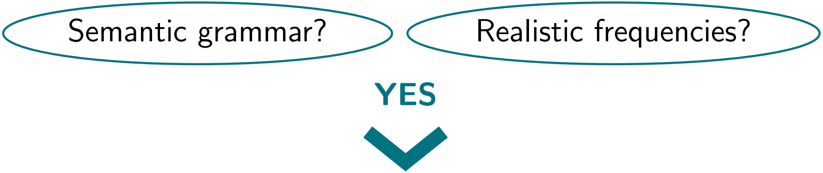
YES

Phrases can be extracted and a translation table construct in a SMT-like way.

Hybrid approaches

Soft integration led by SMT (II)

GF generated corpus



Semantic grammar?

Realistic frequencies?

YES

Phrases can be extracted and a translation table construct in a SMT-like way.

> *!* Many-to-many alignments should be exploited.

Hybrid approaches

Soft integration led by SMT (II)

Ongoing experiments

- 5000 sentences from resource grammar with alignments.
  semantic?
- Many-to-many alignments simulate one-to-many by using multiwords.
- Standard phrase extraction methods can then be used without loosing the power of high quality alignments.
- Probabilities extracted by frequency counts.
  representative?

Hybrid approaches

Soft integration led by GF

Complement with SMT options the GF translation structure.

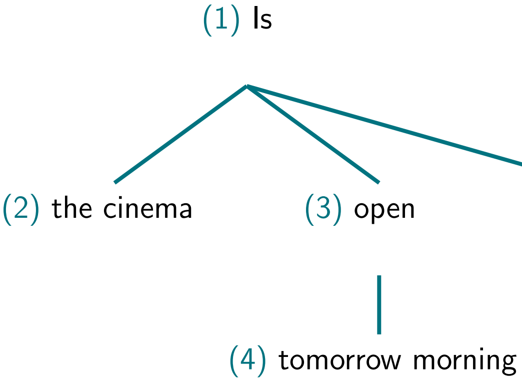
Approach being applied for Spanish-to-Basque with an **RBMT system** (Matxin).

UPC+EHU collaboration.

Applicable to MOLTO?

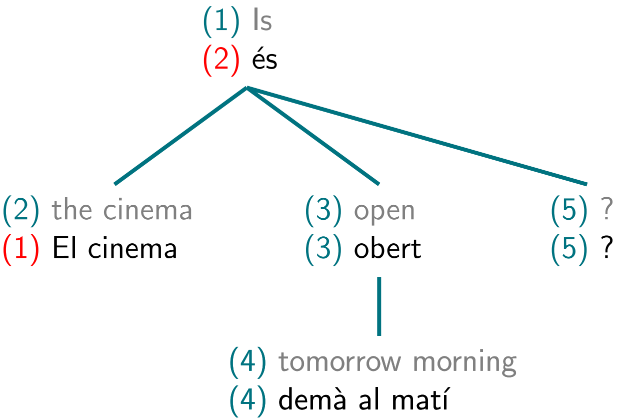
Hybrid approaches

Soft integration led by GF



Hybrid approaches

Soft integration led by GF



(1) Is
(2) és

(2) the cinema
(1) El cinema

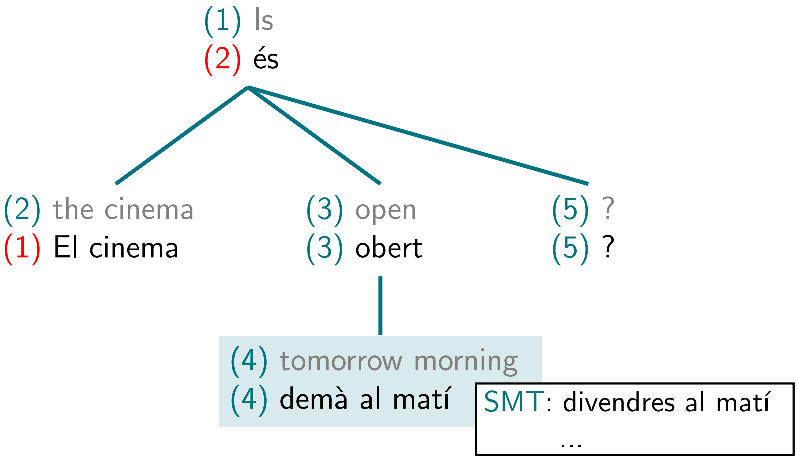
(3) open
(3) obert

(5) ?
(5) ?

(4) tomorrow morning
(4) demà al matí

Hybrid approaches

Soft integration led by GF



(1) Is
(2) és

(2) the cinema
(1) El cinema

(3) open
(3) obert

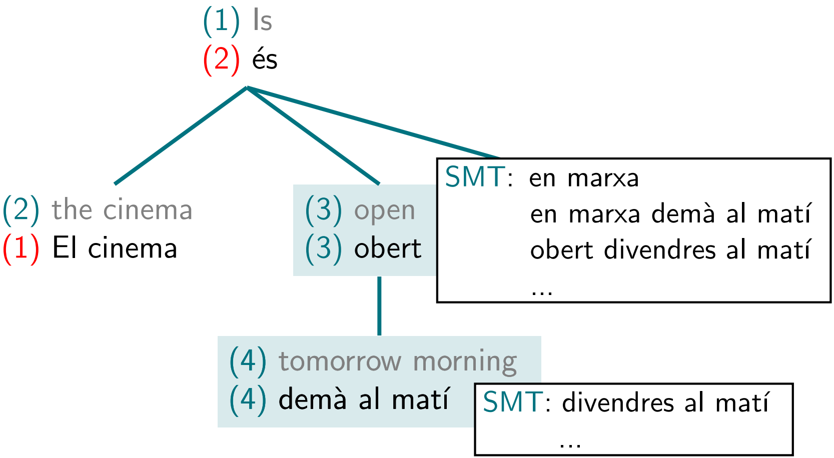
(5) ?
(5) ?

(4) tomorrow morning
(4) demà al matí

SMT: divendres al matí
...

Hybrid approaches

Soft integration led by GF



(1) Is
(2) és

(2) the cinema
(1) El cinema

(3) open
(3) obert

SMT: en marxa
en marxa demà al matí
obert divendres al matí
...

(4) tomorrow morning
(4) demà al matí

SMT: divendres al matí
...

Hybrid approaches

Soft integration led by GF

Comments

- The RB system must parse and translate the input sentence (all!).
- Phrases and segmentation are those given by the RB system.
- Each segment (and up) is sent to a generic SMT to provide more partial translations.
- A second SMT is fed with only the resulting phrases.
- This SMT decoder performs no reordering.

Short term tasks

Todo's and questions to answer

1. Construct (toy?) patents **corpus**. – WP7–
    - Definition, alignment and annotation.

Short term tasks

Todo's and questions to answer

1. Construct (toy?) patents **corpus**. – WP7–
   - Definition, alignment and annotation.

2. Integration of GF **translation table** (TT).
   - Define domain and sets for the subtask.
   - Meaningful probabilities for GF phrases.
   - Joining 3 TTs: too many parameters? having different scores, is it a fair comparison?

Short term tasks

Todo's and questions to answer

3. GF high quality **alignments**.
   - Domain and sets as in number 2.
   - Study the repercussion in SMT.

Short term tasks

Todo's and questions to answer

3. GF high quality **alignments**.
   - Domain and sets as in number 2.
   - Study the repercussion in SMT.

4. Is a Matxin-like **hybrid** viable with GF?
   - Could GF parse a general sentence? Give partial translations?

Short term tasks

Todo's and questions to answer

5. Probabilistic predictions on GF **partial analyses**.
   - Rank or weight ambiguous translations.

Short term tasks

Todo's and questions to answer

5. Probabilistic predictions on GF **partial analyses**.
   - Rank or weight ambiguous translations.

6. GF **grammar** for patents domain.
   - CLEF-IP 2010 data is enough?

Short term tasks

Todo's and questions to answer

5. Probabilistic predictions on GF **partial analyses**.
   - Rank or weight ambiguous translations.

6. GF **grammar** for patents domain.
   - CLEF-IP 2010 data is enough?

Joint work with UGOT: Upcoming internal workshop.

WP5
Statistical and Robust Translation

Cristina España-Bonet
Lluís Màrquez

TALP Research Center

1st MOLTO Project Meeting

Varna, September 9th, 2010