WP5
Statistical and Robust Translation

Lluís Màrquez
Cristina España-Bonet
Universitat Politècnica de Catalunya, TALP Research Center

– 1st year review –

Luxembourg, March 15th, 2011

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**	Probabilistic extension of GF, synthetic corpora for SMT

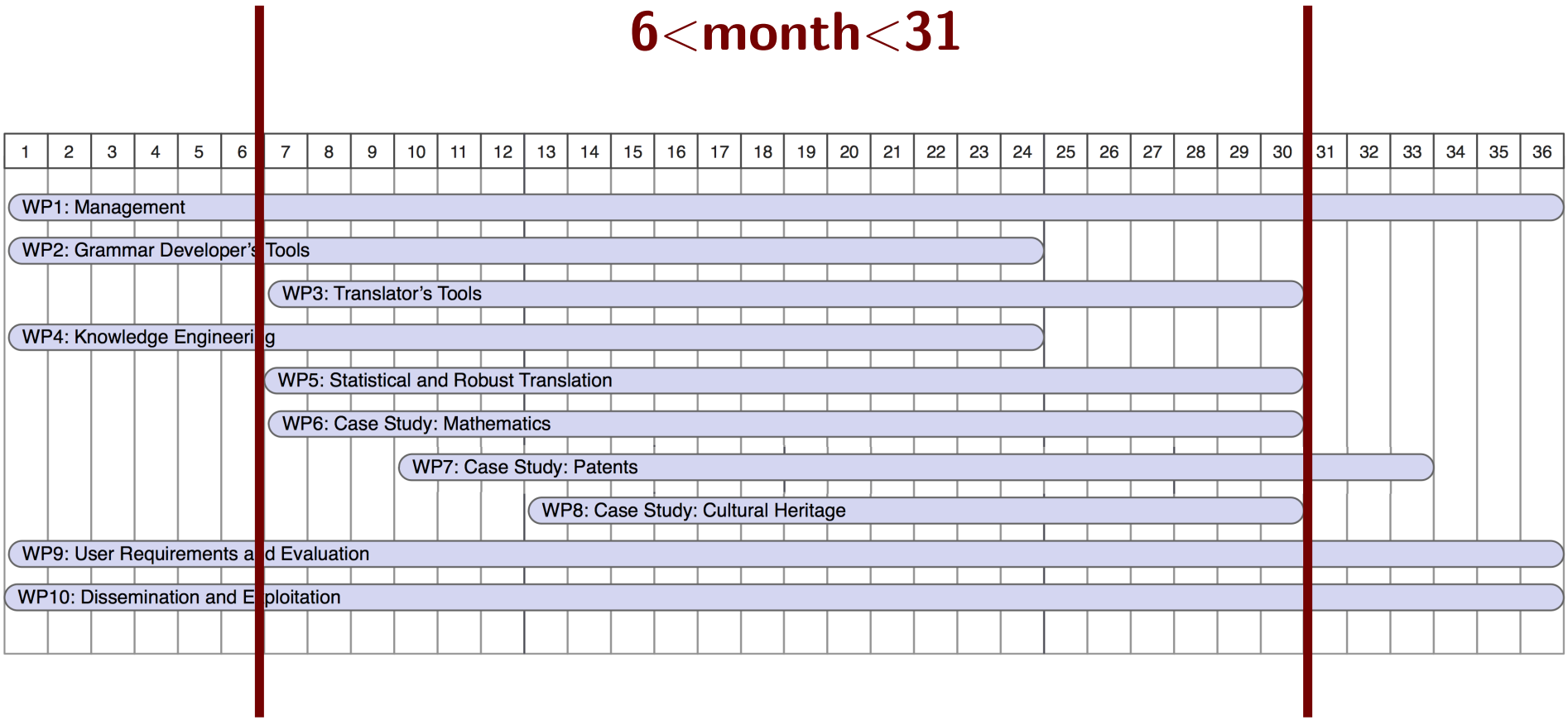
General view

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

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

First year plan, M7-M18

- 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.

Ongoing work

First year plan, status

WP5 is tightly connected to **WP7** (Case of study: Patents).

Consequences:

- 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.

Ongoing work

Work on Baselines

SMT baseline, Standard In-Domain System

- **Corpus**: WP7 selected corpus
- **Language model**: 5-gram interpolated Kneser-Ney discounting, SRILM Toolkit
- **Alignments**: GIZA++ Toolkit
- **Translation model**: Moses package
- **Weights optimization**: MERT against BLEU
- **Decoder**: Moses

Ongoing work

SMT baseline, evaluation

BLEU

	EN2DE	DE2EN	EN2FR	FR2EN	DE2FR	FR2DE
Bing	0.33	0.43	0.43	0.45	0.20	0.24
Google	0.45	0.58	0.53	0.62	0.43	0.39
Domain	**0.58**	**0.65**	**0.62**	**0.70**	**0.56**	**0.53**

Ongoing work

SMT baseline, deep evaluation

	DE2EN			EN2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.52	0.64	**0.72**	0.42	0.51	**0.69**
1-PER	0.66	0.76	**0.82**	0.56	0.64	**0.77**
1-TER	0.59	0.67	**0.76**	0.45	0.53	**0.71**
BLEU	0.43	0.58	**0.65**	0.33	0.45	**0.58**
NIST	8.25	9.67	**10.12**	6.53	8.05	**9.40**
ROUGE-W	0.40	0.48	**0.52**	0.34	0.41	**0.48**
GTM-2	0.30	0.40	**0.47**	0.25	0.32	**0.43**
METEOR-pa	0.60	0.69	**0.74**	0.36	0.45	**0.57**
ULC	0.09	0.29	**0.49**	0.03	0.19	**0.43**

Ongoing work

Two hybridisation approaches: Who leads?

1. Integration led by **SMT**

Make available GF translations to a SMT system.

2. Integration led by **GF**

Complement with SMT options the GF translation structure.

Ongoing work

Two hybridisation approaches: Who leads?

1. Integration led by **SMT**

Make available GF translations to a SMT system.

- If **GF is able to generate Giza-like alignments**, phrases can be extracted in the SMT way and we can combine translation tables.

2. Integration led by **GF**

Complement with SMT options the GF translation structure.

- GF needs **robust and probabilistic parsing** for out of coverage content if has to be applied to open-domain text.

Ongoing work

Hybridisation 1: Giza-like implementation in GF

From many-to-many to one-to-many

```
You want_to_go to the_nearest park
(0)     (1)    (2)     (3)    (4)

Quieres ir al parque mas cercano
(0)    (1)(2)  (3)    (4)    (5)

1-0 1-1 2-2 3-4 3-5 4-3
```

(alignments from Phrasebook grammar)

Ongoing work

Hybridisation 1: experiments

Phrasebook grammar (toy example)

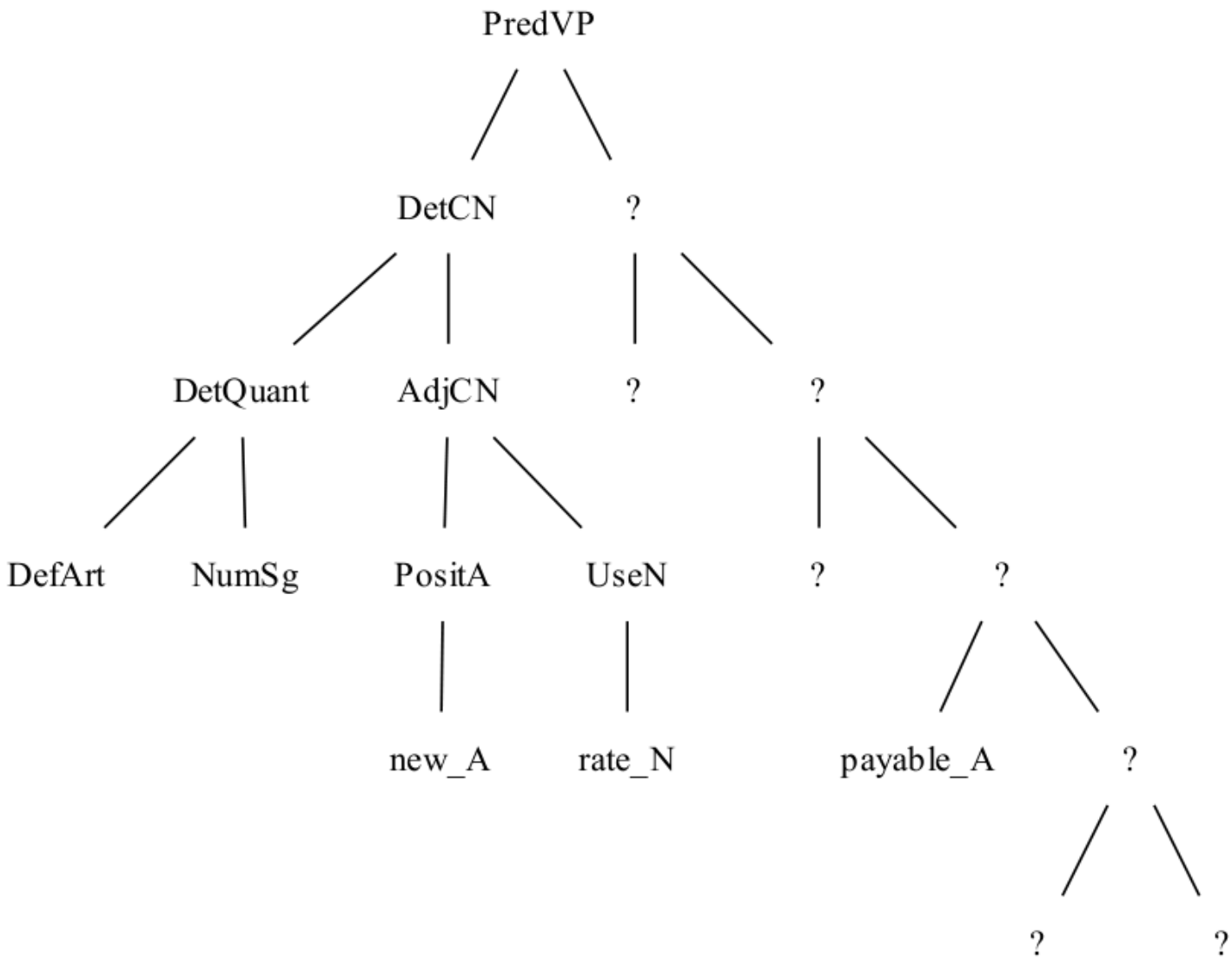
- Syntetic corpus generation
- Parallel corpus with 200 sentences
- Tiny for SMT
- Null intersection with SMT corpora

Patents grammar

- Needed for real experiments (planned within next 6 months)

Ongoing work

Hybridisation 2: robust parsing



Ongoing work

Hybridisation 2: experiments

Chunk parsing. Detect:

- Basic noun phrases i.e. without PP atachement
- Verb phrases without the object
- Prepositions - mark the PP atachements

Experiment on parsing noun phrases from PennTreebank:

- 75% of the phrases were parsed
- Must improve NE and dates

Future work

Overview

Future work

Related to the baselines

- Estimate a **GF baseline** on the test sets defined in WP7.
- **Naïve combination** of GF and SMT as a hybrid baseline.
- **Evaluation** of both systems and comparison with the SMT baseline.

Future work

Related to the hybridisation

- **Hard integration** GF+SMT
  Force fixed GF translations within a SMT system.
- Application of the soft integration **GF+SMT led by SMT** to the patents case, and extension with probabilistic estimations for GF phrases.
- Improvement of the robust parser and implementation of a soft integration **led by GF**.
- A first automatic **evaluation** of the resulting systems.

Dissemination

Overview

Dissemination

WP dissemination

Talks

- *Online Parsing, Type Checking and Advanced Editor for Controlled Languages in GF.* Krasimir Angelov
  MOLTO's second meeting. 8 September, 2010, Varna.

- *Soft integration SMT/GF*
  Cristina España-Bonet and Lluís Màrquez
  MOLTO's internal workshop. 1-5 November, 2010, Chalmers University of Technology, Goteborg.

- *A TAG formalism for Parsing and Translation*
  Xavier Carreras
  MOLTO's internal workshop. 1-5 November, 2010, Chalmers University of Technology, Goteborg.

Dissemination

WP dissemination

Related reports

- *SMatxinT, the Spanish-to-Basque hybrid translator*
  Cristina España-Bonet, Gorka Labaka, Lluís Màrquez and Kepa Serasola
  Internal Report.

WP5
Statistical and Robust Translation

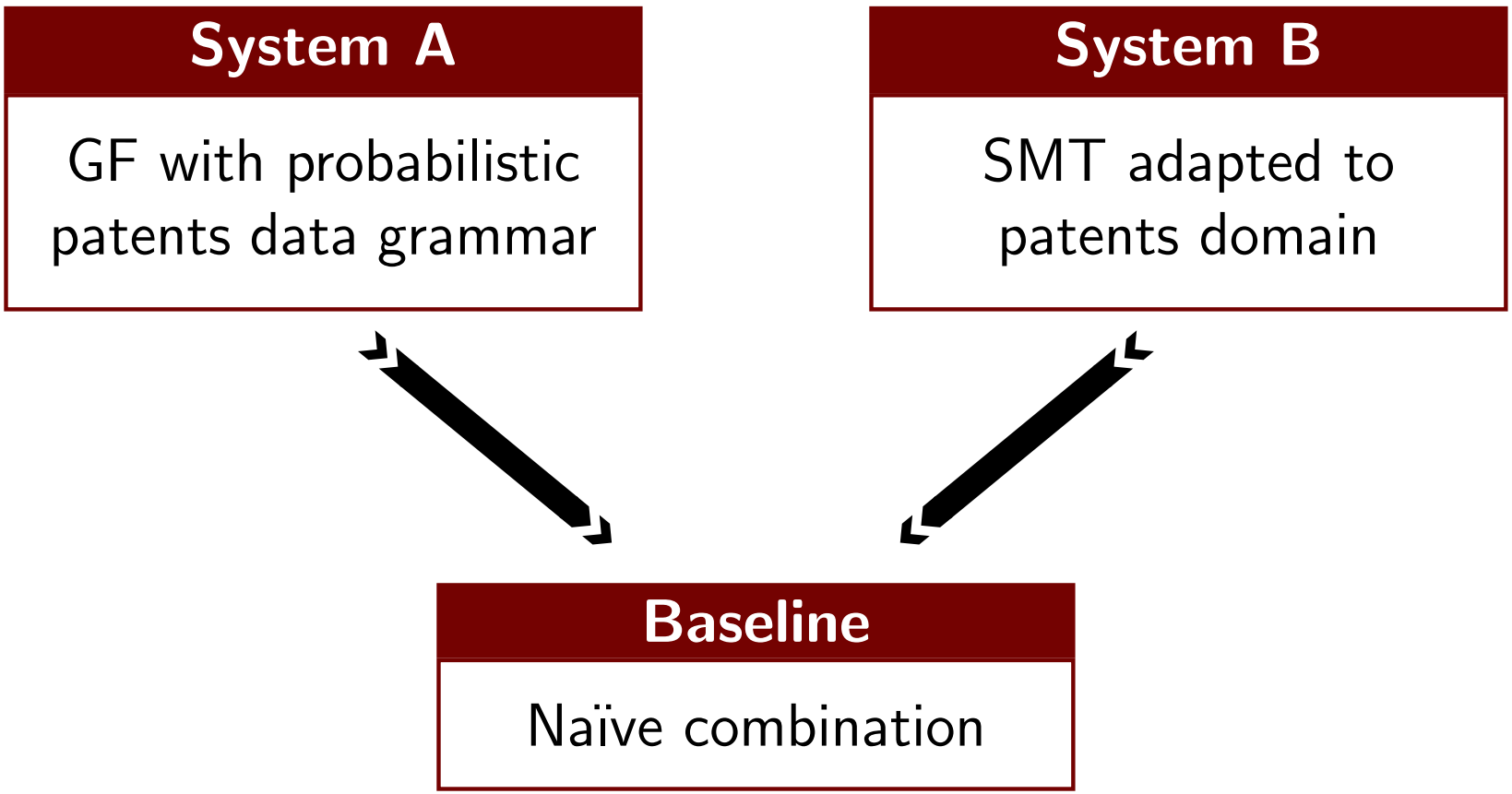
Lluís Màrquez
Cristina España-Bonet
Universitat Politècnica de Catalunya, TALP Research Center

– 1st year review –

Luxembourg, March 15th, 2011

Extra slides

Baseline systems (Ongoing work: System B)



Extra slides

SMT baseline analysis. English-German Translations, scores

	DE2EN			EN2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.52	0.64	**0.72**	0.42	0.51	**0.69**
1-PER	0.66	0.76	**0.82**	0.56	0.64	**0.77**
1-TER	0.59	0.67	**0.76**	0.45	0.53	**0.71**
BLEU	0.43	0.58	**0.65**	0.33	0.45	**0.58**
NIST	8.25	9.67	**10.12**	6.53	8.05	**9.40**
ROUGE-W	0.40	0.48	**0.52**	0.34	0.41	**0.48**
GTM-2	0.30	0.40	**0.47**	0.25	0.32	**0.43**
METEOR-pa	0.60	0.69	**0.74**	0.36	0.45	**0.57**
ULC	0.09	0.29	**0.41**	0.03	0.19	**0.43**

Extra slides

SMT baseline analysis. English-German Translations, examples

Why such good scores?

DE	Verwendung nach Anspruch 23 , worin das molare Verhältnis von Arginin zu Ibuprofen 0,60 : 1 beträgt .
EN	The use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

Extra slides

SMT baseline analysis. English-German Translations, examples

Why such good scores?

DE	Verwendung nach Anspruch 23 , worin das molare Verhältnis von Arginin zu Ibuprofen 0,60 : 1 beträgt .
EN	**The use** of claim 23 , wherein the molar ratio of arginine to ibuprofen **is** 0.60 : 1 .
Domain	The use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .
Google	The **method** of claim 23 , wherein the molar ratio of arginine to ibuprofen 0.60 : 1 **is** .
Bing	The Use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

Extra slides

SMT baseline analysis. English-German Translations, examples

What's wrong?

DE	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
EN	(±)-N-(3-**a**minopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradeceneyloxy)-1-propanaminium **bromide**

Extra slides

SMT baseline analysis. English-German Translations, examples

What's wrong?

DE	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
EN	(±)-N-(3-**a**minopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradeceneyloxy)-1-propanaminium **bromide**
Domain	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
Google	(±)-N-(3-aminopropyl)-N , N-dimethyl-2 , 3-bis (syn-9-tetradecenyloxy) is 1-propanaminiumbromid
Bing	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid

Extra slides

SMT baseline analysis. English-French Translations, scores

	FR2EN			EN2FR		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.54	0.66	**0.78**	0.57	0.63	**0.73**
1-PER	0.71	0.78	**0.86**	0.68	0.75	**0.82**
1-TER	0.59	0.70	**0.80**	0.60	0.66	**0.74**
BLEU	0.45	0.62	**0.70**	0.43	0.53	**0.62**
NIST	8.52	10.01	**10.86**	8.39	9.21	**9.96**
ROUGE-W	0.41	0.50	**0.54**	0.39	0.45	**0.49**
GTM-2	0.32	0.43	**0.53**	0.31	0.36	**0.45**
METEOR-pa	0.61	0.72	**0.77**	0.57	0.65	**0.71**
ULC	0.07	0.28	**0.44**	0.10	0.23	**0.39**

Extra slides

SMT baseline analysis. German-French Translations, scores

	DE2FR			FR2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.42	0.52	**0.76**	0.30	0.43	**0.65**
1-PER	0.58	0.68	**0.77**	0.46	0.59	**0.74**
1-TER	0.47	0.56	**0.68**	0.32	0.46	**0.66**
BLEU	0.29	0.43	**0.56**	0.24	0.39	**0.53**
NIST	6.72	8.21	**9.10**	5.35	7.30	**8.88**
ROUGE-W	0.31	0.38	**0.45**	0.29	0.37	**0.44**
GTM-2	0.24	0.30	**0.41**	0.21	0.28	**0.41**
METEOR-pa	0.45	0.56	**0.64**	0.26	0.39	**0.51**
ULC	0.03	0.22	**0.41**	-0.03	0.19	**0.44**

Extra slides

SMT Systems, general impressions (public systems)

Google

Few OOVs but tokenization problems with compounds.

Bing

Lack of specific vocabulary.

In-domain SMT

Try to solve the problems of the general systems, but still:

- Improve compound detector.
- Fix structures are translated different depending on the vocabulary.

Extra slides

GF baseline (near future work)

GF System

- Composition of **parsing** and **linearisation** via an **abstract syntax** or interlingua

Patents grammar

- **General** structure grammar
- **Compounds** grammar

Extra slides

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Extra slides

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Extra slides

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Missing constituents (verb)

DE	Verwendung nach Anspruch 2, wobei die Menge von Cumarin oder 7-Hydroxycumarin im Medikament 45 mg pro Medikamenten-Einheit **beträgt**.
EN	Use according to claim 2 wherein the amount of coumarin or 7-hydroxycoumarin in the medicament **is** 45 mg pro drug unit.
SMT	The use according to claim 2, wherein the amount of cumarine or 7-Hydroxycumarin in the medicament ϕ 45 mg per Medikamenten-Einheit.

Extra slides

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Reordering problems (verbs & conjunctions)

DE	Verfahren nach Anspruch 20 oder 21, wobei das auf Platin basierende Analogon Cisplatin oder Carboplatin **ist**.
EN	The method of claim 20 or 21, wherein the platin-based analogue **is** cisplatin or carboplatin.
SMT	A method according to claim 20 or 21, wherein the platinum based on analog cisplatin or **is** carboplatin.

Extra slides

Two hybridisation approaches: Who leads?

0. Hard integration

Force fixed GF translations within a SMT system.

1. Soft integration led by **SMT**

Make available GF translations to a SMT system.

2. Soft integration led by **GF**

Complement with SMT options the GF translation structure.

Extra slides

Hybridisation 1: setting

SMT leads translation, RBMT complements

Complement the SMT translation table with RBMT options.

- **GF environment**

  GF alignments for SMT, therefore **language-independent** approach.

  (soon applied to WP7 languages)

Extra slides

Hybridisation 1: SMT vs. GF alignments

GF alignments

- Based on the relation between the concrete syntaxes and the abstract syntax.
- Many-to-many.
- Semantic wrt. abstract syntax.

SMT alignments

- Based on corpus occurrences.
- One-to-many.

Extra slides

Hybridisation 1: experiments (future work)

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` $\mathbf{\log P_{GF}}(e|f)$

Requirements:

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

Extra slides

Hybridisation 2: setting (future work)

GF leads translation, SMT decodes

Complement the GF translation structure with SMT options.

- **GF**

  Nowadays, there is no GF grammar for SMT corpora domains and no SMT corpora for GF grammar domains.

  SMatxinT: Proof of concept.

Extra slides

Hybridisation 2: methodology (future work)

- The GF system must parse and translate the input sentence.
- Phrases and segmentation are those given by the GF system.
- Each segment (and up) is sent to a generic SMT to provide more partial translations.
- A Moses-like decoder is fed with the resulting phrases to search for the highest scored translation.
- This statistical decoder performs no reordering and uses very simple features.

SMatxinT vs. MOLTO

General translator vs. in-domain translator

- With SMatxinT, results are better for **out-of-domain** tests, where the difference between SMT and RBMT systems is less important, but systems (specially SMT) have a lower quality.
- With MOLTO, both systems will be **in-domain**, so they are expected to be high quality. Improvements here will be over already good translations.