WP5
Statistical and Robust Translation

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

–1st year Project Meeting–

Göteborg, March 10th, 2011

WP5

Overview

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

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

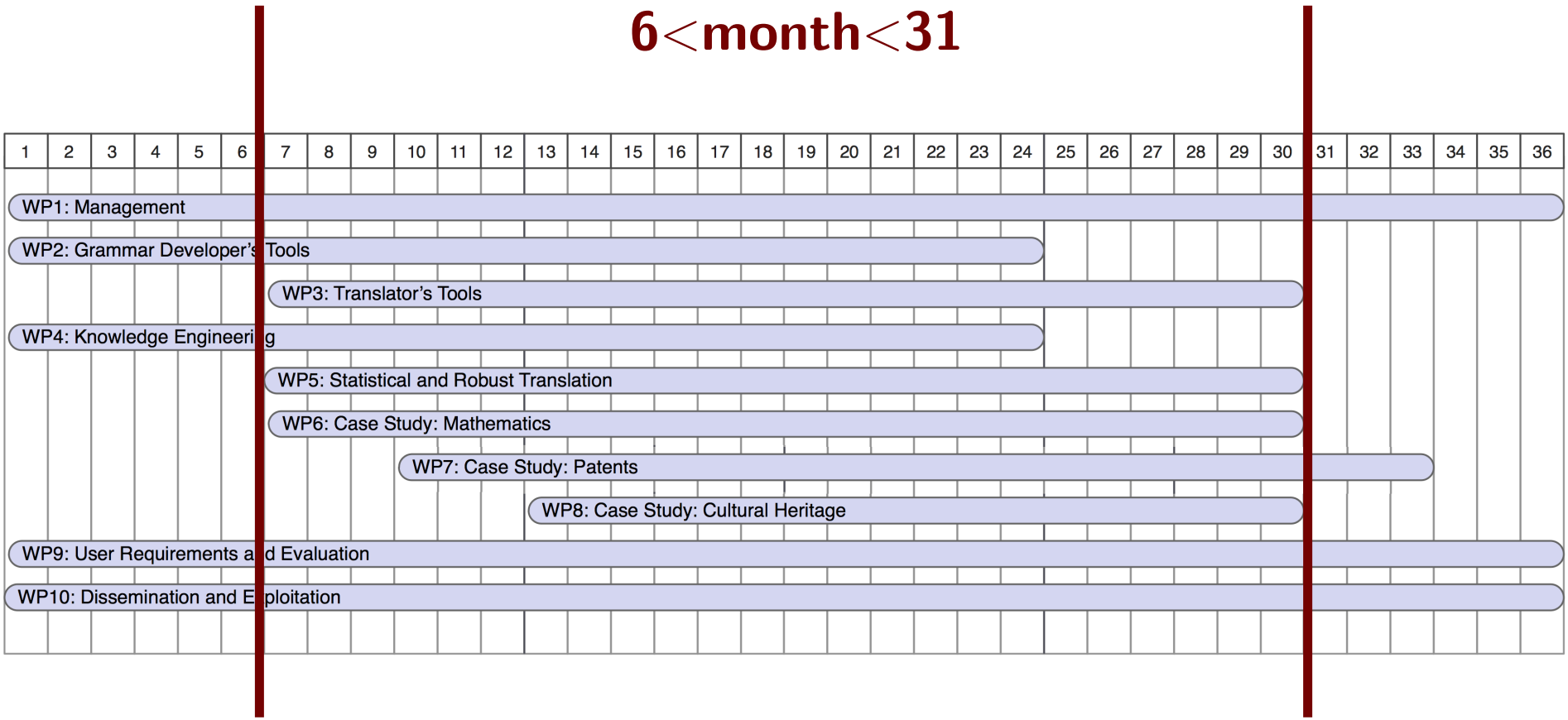
General view

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

1 General view

2 Ongoing work
- Scheduled plan
- Baselines
- Hybrid systems

3 Future work

4 Publications

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

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

Work on hybridisation

SMT leads translation, GF complements

Complement the SMT translation table with GF options.

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

Ongoing work

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 experiments

Phrasebook grammar (toy example)

- Syntetic corpus generation
- Parallel corpus with 200 sentences
- Insignificant for SMT (by 2-3 orders of magnitude!)
- Null intersection with SMT corpora

Patents grammar

- Needed for real experiments

Ongoing work

Robust parsing

Should we include Krasimir's robust parser in WP5?

Future work

Overview

1 General view

2 Ongoing work

3 Future work
- Related to the baselines
- Related to the hybridisation

4 Publications

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.
- Implementation of a soft integration **led by GF**.
- A first automatic **evaluation** of the resulting systems.

Publications

Overview

1 General view

2 Ongoing work

3 Future work

4 Publications

Publications

WP dissemination

Talks

- *Molto Project: Multi Lingual On-Line Translation*
  Cristina España-Bonet.
  Translation Industry among Romanic Languages, València, 2010

- *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.

Publications

WP dissemination

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

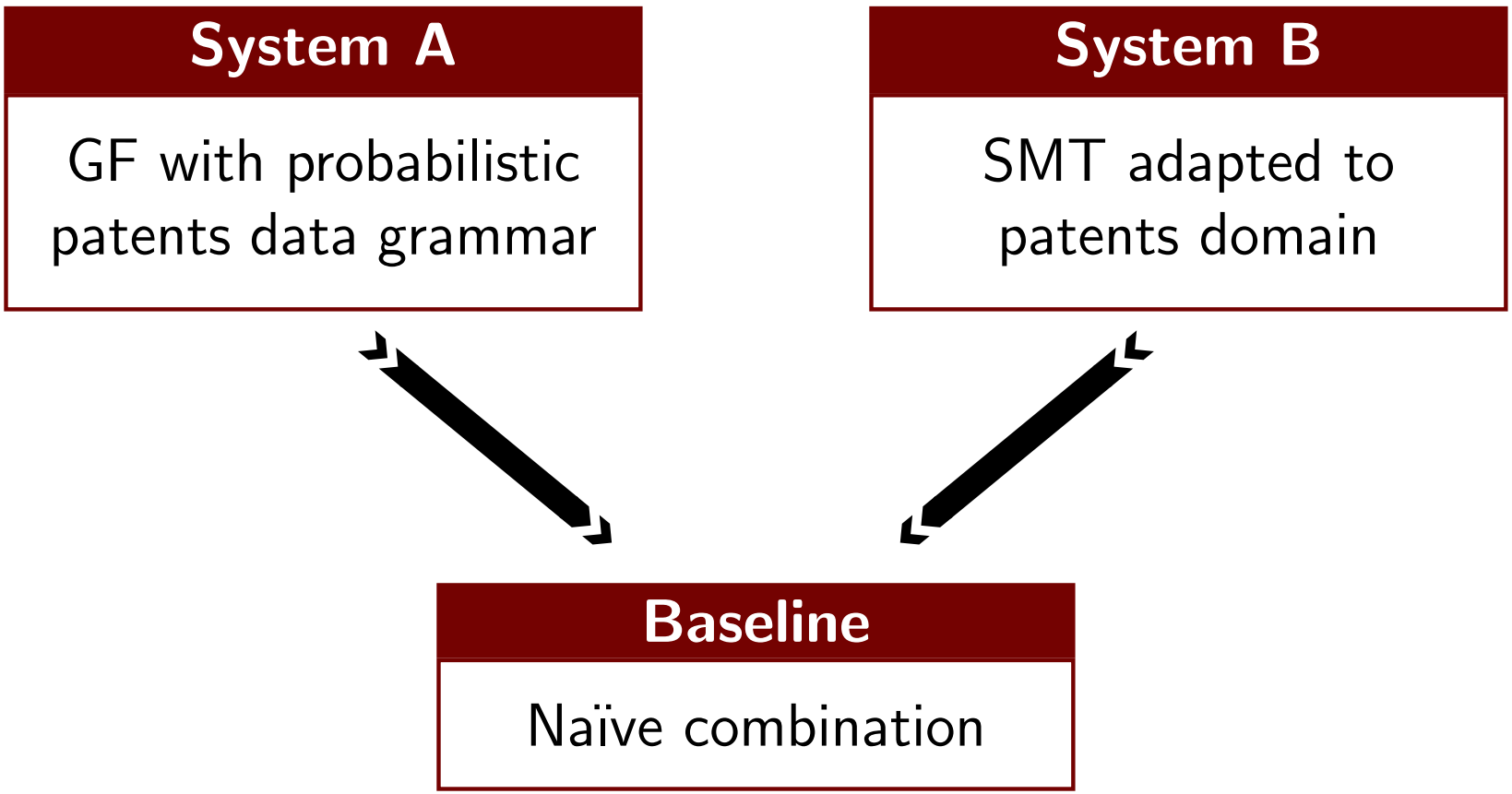
Lluís Màrquez
Cristina España-Bonet
Universitat Politècnica de Catalunya, TALP Research Center

–1st year Project Meeting–

Göteborg, March 10th, 2011

Publications

Hybridisation: Baseline systems



Publications

Baseline, SMT System

Standard In-Domain System

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

Publications

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**

Publications

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 .

Publications

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 .

Publications

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**

Publications

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-aminopropyl)-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

Publications

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**

Publications

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**

Publications

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.

Publications

Baseline, GF System

GF System

- **Parse**
- Apply patents **grammar**
- **Linearise**

Patents grammar

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

Publications

Rule Based MT: Pros and Cons

Pros (as compared to SMT)

- Capture **long distance** relations and reordering.
- Better **grammaticality.**
- (More **robust** to domain changes.)

Cons

- Dependence on the **initial parsing.**
- Lexical transfer **disambiguation.**
- High development **cost** of the grammars and associated resources.

Publications

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Publications

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Publications

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Who leads the hybrid model?

SMT. RBMT is used to enrich the "translation model" of the SMT system (known approach)

RBMT. SMT is used to provide confidence scored translation options to the RBMT target tree (novel)
–addresses cons number 1 and 2 of previous slide–

Publications

An hybrid SMT-RBMT system

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)

Publications

Hybrid SMT-RBMT: GF vs. SMT 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.

Publications

Hybrid SMT-RBMT: Alignment equivalence

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)

Publications

Hybrid SMT-RBMT: Experiments

Phrasebook grammar (toy example)

- Syntetic corpus generation.
- Parallel corpus with 200 sentences.
- Insignificant for SMT (by 2-3 orders of magnitude!).
- Null intersection with SMT corpora.

Patents grammar

- Needed for real experiments.

Publications

Hybrid SMT-RBMT: Experiments on combination

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.

Publications

An hybrid RBMT-SMT system: SMatxinT

RBMT leads translation, SMT decodes

Complement the RBMT translation structure with SMT options.

- **SMatxinT**

  Approach being applied for **Basque-to-Spanish** with the RBMT system Matxin.

  `OpenMT-2 Spanish Research Project`
  `UPC+EHU collaboration`

Publications

An hybrid RBMT-SMT system: SMatxinT, methodology

- The RBMT system must parse and translate the input sentence.
- Phrases and segmentation are those given by the RBMT 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.

Publications

An hybrid RBMT-SMT system: SMatxinT, comments

Current results

- Large difference between in-domain and out-of-domain scenarios.
- Results are at most close to SMT system.
- Oracles show large room for improvement.
- RBMT phrases are underused.
- Current features are not distinctive enough.

Publications

SMatxinT in relation with MOLTO

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 quallity.

Matxin vs. GF

General grammar vs. in-domain grammar

With MOLTO both systems will be in-domain, so they are expected to be high quality. Improvements here will be over already good translations.

Publications

Statistical extension of GF grammar

Learning GF grammars

Abstract syntax	Like She He	Grammarian
English example	she likes him	Grammarian
German translation	er gefällt ihr	**SMT**
Resource tree	mkCl he_{Pron} $\text{gefallen}_{\text{V2}}$ she_{Pron}	GF parser
Syntax rule	Like x y = mkCl y $\text{gefallen}_{\text{V2}}$ x	Variables renamed

- SMT of short and frequent sentences is good

Publications

Statistical extension of GF grammar, application

- Applied to the **Phrasebook grammar**
- **Languages**: Danish, Dutch, German, Norwegian
- Phrasebook **demo**:
  `http://www.molto-project.eu/demo/phrasebook`