Statistical and robust translation in MOLTO

Cristina España-Bonet, Ramona Enache,
Aarne Ranta and Lluís Màrquez

Universitat Politècnica de Catalunya, University of Gothenburg

– Second year project meeting –

Zurich, March 7th, 2012

Overview

1. Introduction

2. Translation Systems
   - SMT
   - GF
   - Hybrid

3. Conclusions

Introduction

High quality translation

MOLTO Multilingual Online Translation

Non multa, sed multum not quantity but quality

MOLTO aims at high quality translation for concrete domains.

Robustness and domain widening are achived by **SMT** components, still working on a quasi-open domain with a controlled language: **Patents**.

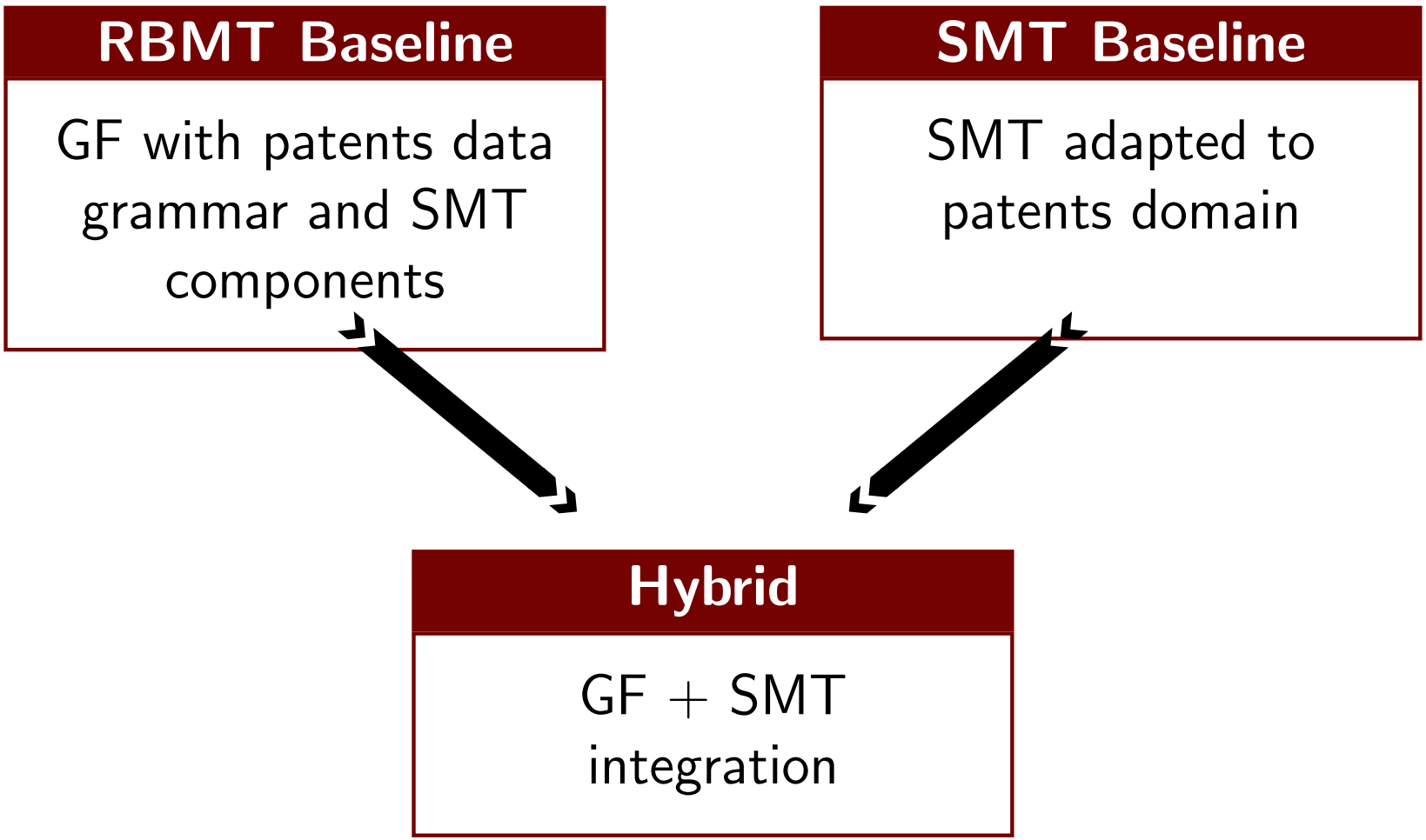
Introduction

Translation System

1. Resources
   - Parallel corpus
   - Grammar
2. Translation engine
   - Statistical, `SMT`
   - Rule based, `GF`
   - Hybrid, `GF+SMT`

Introduction

Engines



Overview

1. Introduction

2. Translation Systems
   - SMT
   - GF
   - Hybrid

3. Conclusions

Translation Systems

What's next...

SMT

Translation Systems

SMT system

SMT baseline, 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
- **Evaluation**: Asiya

Translation Systems

Corpus figures

Tokenized parallel corpus in the chemical domain:

SET	Segments	EN tok	DE tok	FR tok
Training	279,282	7,954,491	7,346,319	8,906,379
Development	993	29,253	26,796	33,825
Test	1,008	31,239	28,225	35,263

Obtained from patent documents with translated claims and IPC classification **A61P** (*Specific therapeutic activity of chemical compounds or medical preparations*).

Translation Systems

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

```
- The use according to claim 7, wherein said cancer diseases comprise
bladder, lung, mamma, melanoma and prostate carcinomas.

- A compound according to claim 1 wherein it is
(2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.

- The pharmaceutical composition according to claim 1 or 2, wherein said
platinum anticancer agent is selected from at least one of the complexes
having structures of:  **IMAGE**.
```

Translation Systems

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- **The use according to claim 7, wherein** `said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.`
- **A compound according to claim 1 wherein** `it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.`
- `The pharmaceutical` **composition according to claim 1 or 2, wherein said** `platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.`

Translation Systems

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise **bladder, lung, mamma, melanoma and prostate carcinomas**.
- A compound according to claim 1 wherein it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.
- The pharmaceutical composition according to claim 1 or 2, wherein said **platinum anticancer agent** is selected from at least one of the complexes having structures of: **IMAGE**.

Translation Systems

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.
- A compound according to claim 1 wherein it is **(2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide**.
- The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Translation Systems

Language domain and genre

Claims have also **long sentences** and **missing information**.

Excerpt 2

- Use of compounds of formula I **IMAGE** wherein R1 signifies substituted C1-C4-alkylene, whereby the substituents are selected from the group comprising unsubstituted aryloxy or aryloxy mono- to penta-substituted by R5, and unsubstituted pyridyloxy or pyridyloxy mono- to tetra-substituted by R5, whereby the substituents may be the same as one another or different if the number thereof is greater than 1; R2 signifies unsubstituted phenyl or phenyl mono- to penta-substituted by R5, or unsubstituted pyridyl or pyridyl mono- to tetra-substituted by R5; R3 is methyl; R4 signifies hydrogen, C1-C6-alkyl or halogen-C1-C6-alkyl; R5 signifies C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy, C2-C6-alkenyl, halogen-C2-C6-alkenyl, C2-C6-alkinyl, halogen-C2-C6-alkinyl, C3-C8-cycloalkyl, C1-C6-alkylcarbonyl, halogen-C1-C6-alkylcarbonyl, C1-C6-alkoxycarbonyl, halogen-C1-C6-alkoxycarbonyl, C1-C6-alkylsulfonyl, C1-C6-alkylsulfinyl, halogen, cyano or nitro; A signifies C(R6)(R7), CH=CH or C=C; R6 and R7 either, i ndependently of one another, signify hydrogen, halogen, C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy or C3-C6-cycloalkyl; or together signify C2-C6-alkylene; R8 and R9 are hydogen; m and n, independently...of one other, are 0 or 1; and optionally enantiomers thereof, with the proviso that if m is 0 then R1 is retained; in the preparation of a pharmaceutical composition for the control of endoparasitic helminths in warm-blooded productive livestock and domestic animals.

Translation Systems

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**

Translation Systems

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**

Translation Systems

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 .

Translation Systems

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 .

Translation Systems

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**

Translation Systems

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

Translation Systems

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**

Translation Systems

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**

Translation Systems

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

Translation Systems

What's next...

GF

Translation Systems

GF for patents

GF System

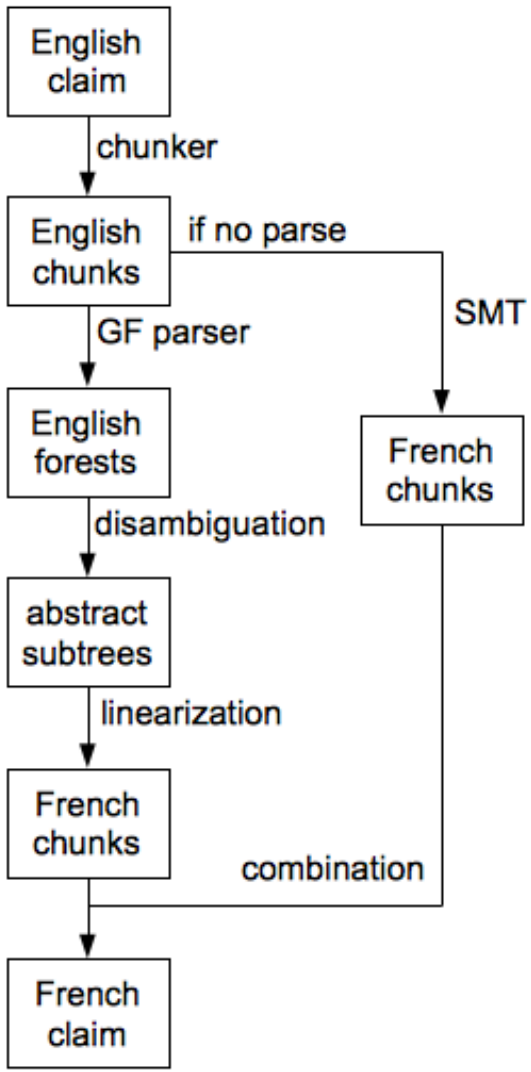
- **Parse**
- Apply patents **grammar**
- **Linearise**

Patents grammar

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

Translation Systems

GF with a patents data grammar



English-to-French patent translator

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**

Chunking

- 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

Translation Systems

GF with a patents data grammar

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

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

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**

HYBRIDS

Translation Systems

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Translation Systems

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Translation Systems

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.

Translation Systems

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.

Translation Systems

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Gender and number agreement

EN	A use according to claim 3 , wherein the **separate** medicament is administered at the same time as...
FR	Utilisation selon la revendication 3, dans laquelle le médicament **séparé** est administré en même temps que...
SMT	Utilisation selon la revendication 3, dans laquelle le médicament **séparée** est administré en même temps que...

Translation Systems

Two hybridisation approaches: Who leads?

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.

3. GF is complemented with SMT translations

Complement the GF translation structure with SMT options.

! The GF baseline system is already hybrid with SMT:

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**

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

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)

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%)*

Example

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

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 .**

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> **.**

2. SMT adds GF translations

GF translations are phrases with a probability within a SMT system.

Characteristics:

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

Example

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

Translation Systems

2. A hybrid SMT-GF system (SI)

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

```
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

Translation Systems

Hybrid results

English-to-French 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

Translation Systems

Hybrid results

English-to-French 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

Example translation

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**...

Example translation

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**...

Conclusions

1. Introduction

2. Translation Systems
   - SMT
   - GF
   - Hybrid

3. Conclusions

Conclusions

Hybrid translation of patents

Concept

Molto aims at increasing the **robustness and coverage** of translation in quasi-open domains by combining GF with SMT.

Conclusions

Hybrid translation of patents

Concept

Molto aims at increasing the **robustness and coverage** of translation in quasi-open domains by combining GF with SMT.

Hybrid engines have been developed to join the coverage of SMT with the precision of GF.

Conclusions

Hybrid translation of patents

Concept

Molto aims at increasing the **robustness and coverage** of translation in quasi-open domains by combining GF with SMT.

Hybrid engines have been developed to join the coverage of SMT with the precision of GF.

GF is expected to provide **grammatically correct** phrases to SMT translations.

Conclusions

Hybrid translation of patents

Observations

The best performance is obtained with the **soft integration** (SMT *can* use GF translations).

Conclusions

Hybrid translation of patents

Observations

The best performance is obtained with the **soft integration** (SMT *can* use GF translations).

GF still has a **low effect** on the final translation.

Conclusions

Hybrid translation of patents

Observations

The best performance is obtained with the **soft integration** (SMT *can* use GF translations).

GF still has a **low effect** on the final translation.

Automatic metrics may not be able to capture the gain in fluency (need for **manual evaluation**).

Conclusions

Hybrid translation of patents

Future work

Tackle **more issues** with the GF grammar (sentence level reordering, compounds, etc.).

Conclusions

Hybrid translation of patents

Future work

Tackle **more issues** with the GF grammar (sentence level reordering, compounds, etc.).

Alternative GF system with SMT translation before parsing.

Conclusions

Hybrid translation of patents

Future work

Tackle **more issues** with the GF grammar (sentence level reordering, compounds, etc.).

Alternative GF system with SMT translation before parsing.

New ways of **phrase combination** (normalise probabilities).

Conclusions

Hybrid translation of patents

Future work

Tackle **more issues** with the GF grammar (sentence level reordering, compounds, etc.).

Alternative GF system with SMT translation before parsing.

New ways of **phrase combination** (normalise probabilities).

Extend the system to deal with **German** translation.

Vielen Dank!

Gràcies!

Thank you!

Statistical and robust translation in MOLTO

Cristina España-Bonet, Ramona Enache,
Aarne Ranta and Lluís Màrquez

Universitat Politècnica de Catalunya, University of Gothenburg

– Second year project meeting –

Zurich, March 7th, 2012

Vielen Dank!

A Patent document

Patent document, **IPC** classification.


```
−<patent-document ucid="EP-1738753-B1" country="EP" doc-number="1738753" kind="B1" lang="EN" date="20080423" family-id="37453347"
  date-produced="20100220" status="new">
  −<bibliographic-data>
    −<publication-reference fvid="88724218" ucid="EP-1738753-B1" status="new">
      −<document-id status="new" format="original">
          <country status="new">EP</country>
          <doc-number>1738753</doc-number>
          <kind>B1</kind>
          <date>20080423</date>
          <lang>EN</lang>
        </document-id>
      </publication-reference>
    +<application-reference mxw-id="PAPP77683688" ucid="EP-06017469-A" load-source="docdb" status="new" is-representative="NO"></application-
      reference>
    +<priority-claims status="new"></priority-claims>
    +<dates-of-public-availability status="new"></dates-of-public-availability>
    −<technical-data status="new">
      −<classifications-ipcr>
          <classification-ipcr mxw-id="PCL624787575" load-source="docdb" status="new">A61K 31/135 20060101C I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624787849" load-source="docdb" status="new">A61P 3/04 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624795950" load-source="docdb" status="new">A61K 31/135 20060101A I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624799973" load-source="docdb" status="new">A61P 25/20 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624806558" load-source="docdb" status="new">A61K 31/137 20060101CFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624810330" load-source="docdb" status="new">A61K 31/137 20060101AFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624820189" load-source="docdb" status="new">A61P 3/00 20060101CLI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624827390" load-source="docdb" status="new">A61P 25/00 20060101ALI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624828549" load-source="docdb" status="new">A61P 25/00 20060101CLI20071018BHEP </classification-ipcr>
```

Vielen Dank!

A Patent document

Description, **claims**.

```
      <u style="single">Obesity Reduction Test Results</u>
    </b>
  </heading>
 −<p num="p0023">
    The venlafaxine group showed consistent statistically significant mean weight decreases and mean percent decreases from baseline beginning at week 1. Overall, the mean decrease in body weight for the venlafaxine group at week 10 was 7.5 lb with a mean percent decrease from baseline of 3.6%. In contrast, the mean decrease in body weight for the placebo group at week 10 was 1.3 lb with a mean percent decrease from baseline of 0.7%. The body mass index evaluation for the venlafaxine also showed a pattern of decreases similar to that of the weight decreases.
  </p>
 </description>
−<claims mxw-id="PCLM12825865" lang="DE" load-source="patent-office" status="new">
 −<claim id="c-de-01-0001" num="0001">
  −<claim-text>
    Verwendung einer Verbindung mit der Formel
   +<chemistry id="chem0006" num="0006"></chemistry>
    in der A eine Komponente der Formel
   +<chemistry id="chem0007" num="0007"></chemistry>
    ist, wobei
    <br/>
    die gestrichelte Linie eine optionale Unsättigung darstellt;
   −<claim-text>
     R
     <sub>1</sub>
     Wasserstoff oder Alkyl mit 1 bis 6 Kohlenstoffatomen ist;
    </claim-text>
   −<claim-text>
     R
     <sub>2</sub>
```

Vielen Dank!

Rule Based MT Systems

- Transfer style translation
- Several sequential steps:
  - Parse input sentence
  - Apply structural and lexical transfer rules
  - Generate output text in the target language
- Transfer grammar: one per language pair
- Parser and generator: one per language

Vielen Dank!

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.

Vielen Dank!

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Vielen Dank!

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Vielen Dank!

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Who leads the hybrid model?

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

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