SMT Applied to the Patent Domain. Hybridisation with GF and RBMT Paradigms.

Cristina España-Bonet and Lluís Màrquez

Universitat Politècnica de Catalunya, TALP Research Center

– First year project meeting –

Göteborg, March 9th, 2011

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**.

Overview

1. Case of Study: Patents

2. Hybridisation
   - Baseline systems
   - Hybridisation techniques
   - Hybridisation examples

3. Conclusions

Case of Study: Patents

Corpus

CLEF-IP 2010 Collection

Extract of the MAREC dataset, containing over 2.6 million patent documents pertaining to 1.3 milion patents from the EPO with some content in English, German and French.

Case of Study: Patents

Parallel corpus selection

- Patent documents with **translated claims**
  (not all of them!)
- IPC classification **A61P**
  Specific therapeutic activity of chemical compounds or medical preparations.

Case of Study: Patents

Parallel corpus selection

- Patent documents with **translated claims**
  (not all of them!)
- IPC classification **A61P**
  Specific therapeutic activity of chemical compounds or medical preparations.

56,000 patents out of 1.3 million fulfill these demands.

(279,282 aligned parallel fragments)

Case of Study: Patents

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**.
```

Case of Study: Patents

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**.`

Case of Study: Patents

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**.

Case of Study: Patents

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**.

Case of Study: Patents

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.

Case of Study: Patents

Chemical domain issues

The main issue is the **treatment of chemical compounds**.

- **Compound detector**
  Based on affix detection.
- **Compound tokenizer**
  Based on the detector and a regular tokenizer.
- **Compound translator**
  Two separate approaches: SMT and GF.

Case of Study: Patents

Compound tokenizer (non-tokenizer!)

Regular tokenizer

8-difluoro-2- [3-fluoro-4 - [(L-lysyl) amino] phenyl] -7-methyl-4H-1-benzopyran-4-one

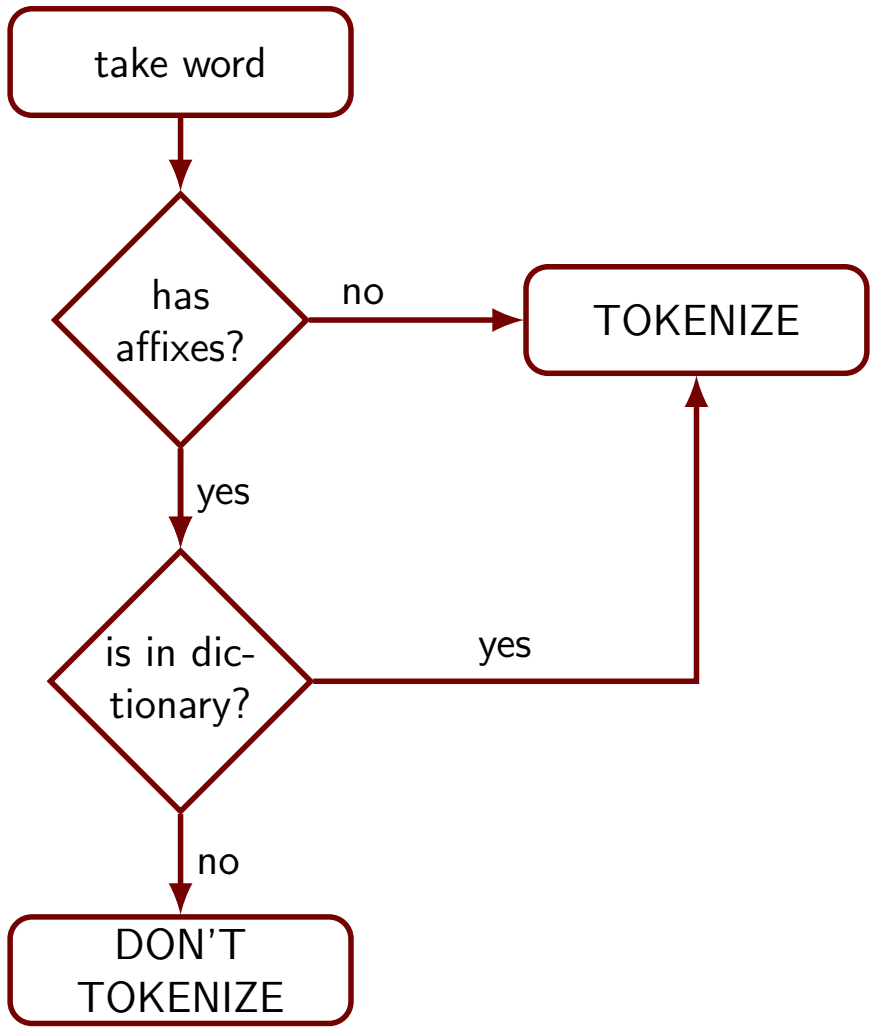
- Parenthesis and square brackets are separated.
- Punctuation is separated.

Desired tokenizer

8-difluoro-2-[3-fluoro-4-[(L-lysyl)amino]phenyl]-7-methyl-4H-1-benzopyran-4-one

Case of Study: Patents

Compound tokenizer (non-tokenizer!)



Case of Study: Patents

Compound tokenizer (non-tokenizer!)

Elements that appear in the **list of affixes**

Prefixes `Meth-, Eth-, Prop-, Pentadec-, imido-, selenocarboxy-, hydroxy-, Propion-, Arachid-...`

Sufixes `-ol, -one, -al, -aldehyde, -oic, -oate, -oxy, -sulfonic, -nitrile, -amine, -isocyanide...`

(English & German: 142 elements, French: 148 elements)

Case of Study: Patents

Compound tokenizer (non-tokenizer!)

Elements that appear in the **list of affixes**

Prefixes `Meth-, Eth-, Prop-, Pentadec-, imido-, selenocarboxy-, hydroxy-, Propion-, Arachid-...`

Sufixes `-ol, -one, -al, -aldehyde, -oic, -oate, -oxy, -sulfonic, -nitrile, -amine, -isocyanide...`

(English & German: 142 elements, French: 148 elements)

Need to check against a **dictionary** (English).

Case of Study: Patents

Compound detection from the tokenizer

The method works better as a tokenizer than as a compound detector, it beds for **high recall** instead of precision.

Actual missclassifications:

- Proper names: `Hôpit`**al**
- Words which are not in the dictionary: `Extracorpore`**al**
- Groups: `-internation`**al**
- Typos: `compar`**oate**

Case of Study: Patents

Compound detection from the tokenizer

The method works better as a tokenizer than as a compound detector, it beds for **high recall** instead of precision.

Actual missclassifications:

- Proper names: `Hôpit`**al**
- Words which are not in the dictionary: `Extracorpore`**al**
- Groups: `-internation`**al**
- Typos: `compar`**oate**

103,272 (compounds + noise)

Case of Study: Patents

Corpus

Final **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

IPC A61P

Overview

1. Case of Study: Patents

2. Hybridisation
   - Baseline systems
   - Hybridisation techniques
   - Hybridisation examples

3. Conclusions

Hybridisation

Translation System

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

Hybridisation

Proposed baselines

RBMT Baseline

GF with probabilistic patents data grammar

SMT Baseline

SMT adapted to patents domain

Hybrid Baseline

Naïve combination

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

Hybridisation

Baseline, SMT System

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**

Hybridisation

Baseline, SMT System

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**

1-TER

	EN2DE	DE2EN	EN2FR	FR2EN	DE2FR	FR2DE
Bing	0.45	0.59	0.60	0.59	0.47	0.32
Google	0.53	0.67	0.66	0.70	0.56	0.46
Domain	**0.71**	**0.76**	**0.74**	**0.80**	**0.68**	**0.66**

Hybridisation

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.

Hybridisation

Baseline, GF System

GF System

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

Patents grammar

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

Hybridisation

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Hybridisation

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Hybridisation

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.

Hybridisation

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.

Hybridisation

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.

SMT leads translation, GF complements

Complement the SMT translation table with GF options.

- **GF environment**

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

  (soon applied to WP7 languages)

Hybridisation

Hybrid SMT-GF: 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

Hybrid SMT-GF: 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)

Hybridisation

Hybrid SMT-GF: 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

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.

RBMT leads translation, SMT decodes

Complement the RBMT translation structure with SMT options.

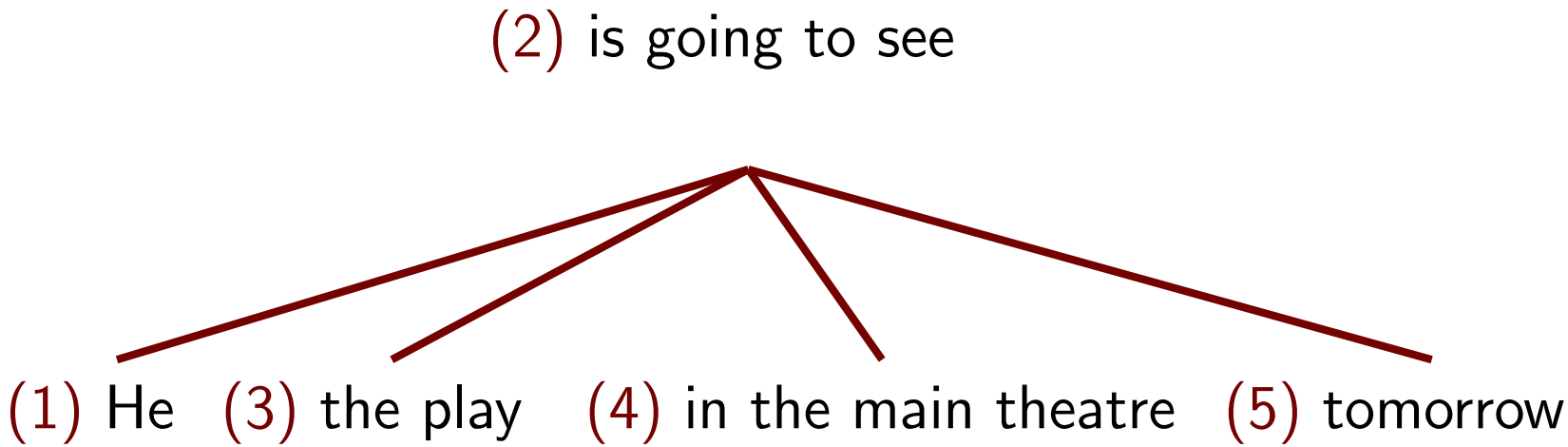
- **SMatxinT**

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

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

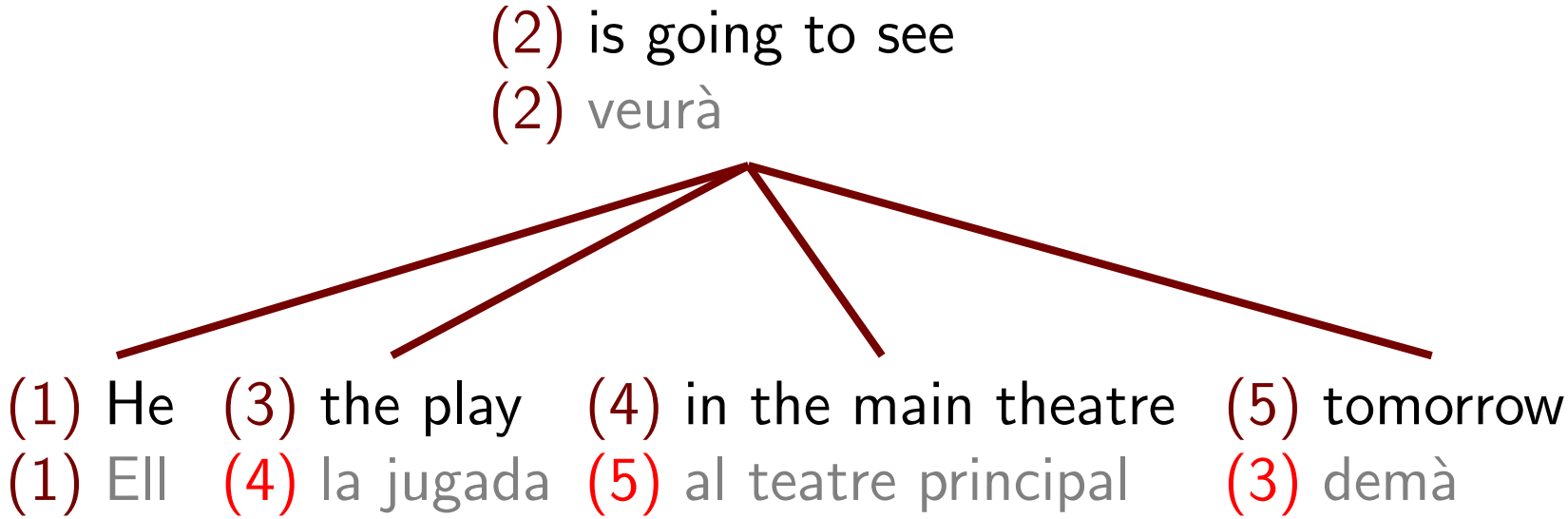
Hybridisation

SMatxinT: Parse tree



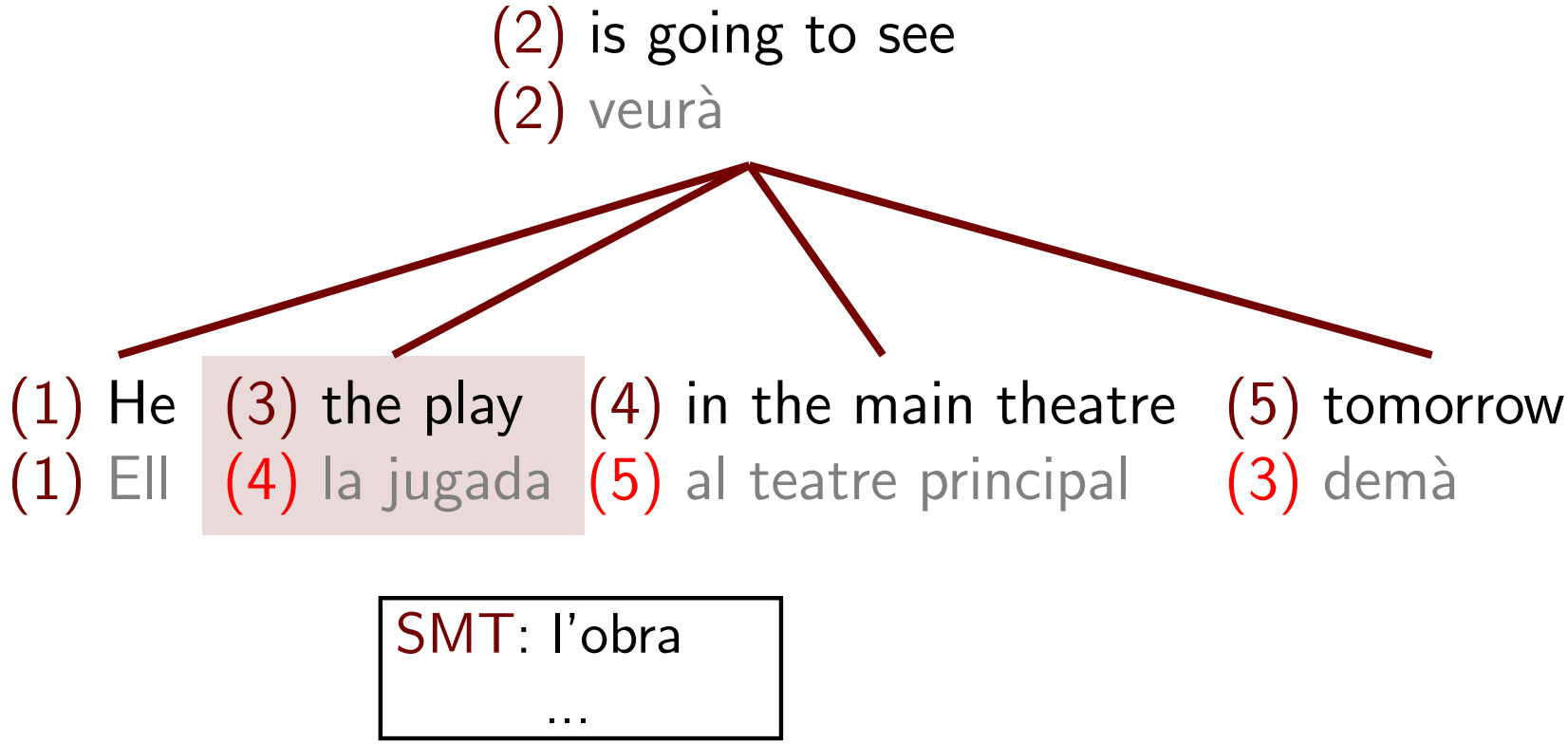
Hybridisation

SMatxinT: Parse tree



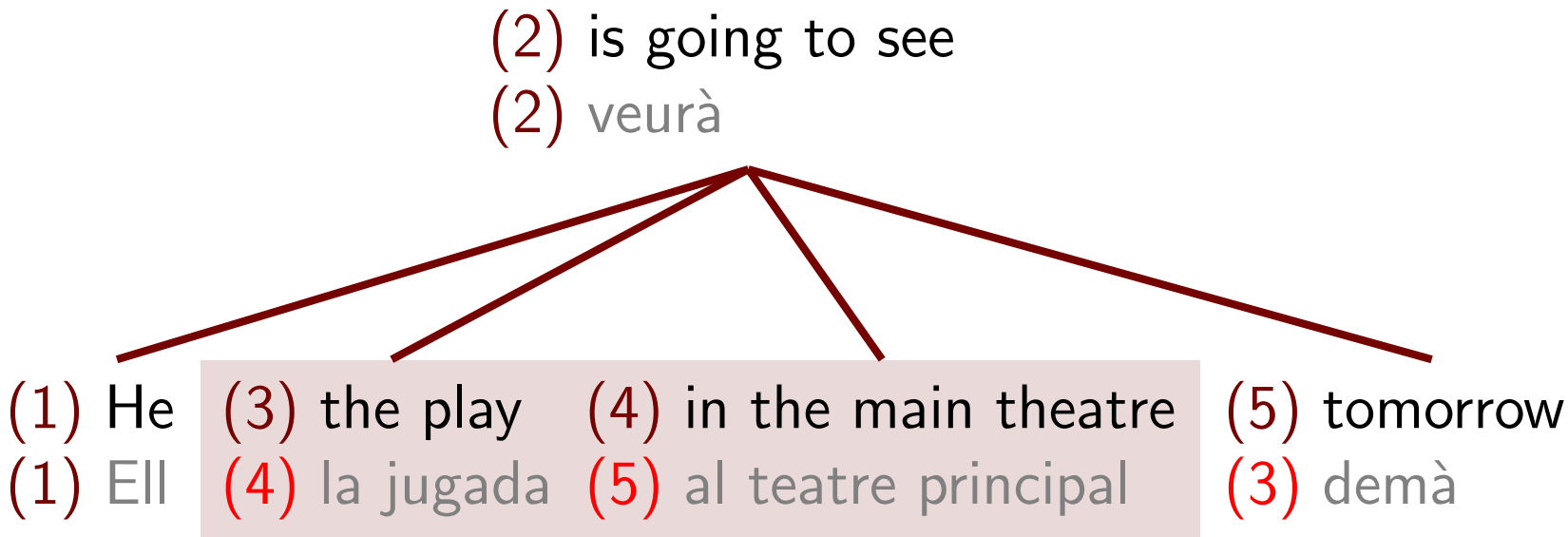
Hybridisation

SMatxinT: Parse tree



Hybridisation

SMatxinT: Parse tree



SMT: l'obra

SMT: l'obra
l'obra al cinema principal
l'obra al teatre principal
...

Hybridisation

SMatxinT: Monotonous decoding

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**

Hybridisation

SMatxinT: Monotonous decoding

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al cinema principal

Hybridisation

SMatxinT: Monotonous decoding

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al cinema principal
Ell	veurà	demà	l'obra	al teatre principal
ϕ	mirarà		la jugada	al cinema principal
	...		...	al teatre del centre

Hybridisation

SMatxinT: Monotonous decoding

(1)	**(2)**	**(5)**	**(3)**	**(4)**
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al cinema principal
Ell	veurà	demà	l'obra	al teatre principal
ϕ	mirarà		la jugada	al cinema principal
	...		...	al teatre del centre
			l'obra al cinema del centre	
			l'obra al teatre principal	
			...	

Hybridisation

SMatxinT: Monotonous decoding

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al cinema principal
Ell ϕ	veurà mirarà ...	demà	l'obra la jugada ...	al teatre principal al cinema principal al teatre del centre
			l'obra al cinema del centre l'obra al teatre principal ...	

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

Hybridisation

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.

Hybridisation

SMatxinT: Oracle translation

Out-of-domain test set (2 refs.)

	RBMT	**SMT**	**Oracle**
BLEU	7.23	7.90	13.64
TER	83.15	79.15	74.81

Large room for improvement!

Hybridisation

SMatxinT: Hybrid translation

	RBMT	**SMT**	**Hybrid**	**Oracle**
BLEU	7.23	7.90	**7.75**	13.64
TER	83.15	79.15	**80.50**	74.81

	Hybrid		**Oracle**	
Procedence	**# chunks**	**% chunks**	**# chunks**	**% chunks**
SMT	2920	**60.2**	3792	**42.6**
RBMT	232	**4.8**	1724	**19.4**
BOTH	1696	35.0	3381	38.0
Total	*4848*	*100*	*8897*	*100*

Example

SPA	El Tour de Flandes se disputa este domingo, y por tanto, el belga Tom Boonen serà el líder del equipo.
EUS	Flandeseko Tourra igande honetan lehiatuko da, eta beraz, Tom Boonen belgikarra izango da taldeko liderra.
RBMT	Tour De Flandes igande honetan eztabaidatzen da, eta beraz, Tom Boonen belgikarra taldearen liderra izango da.
SMT	Flandesko itzulia jokatuko igande honetan, eta beraz, belgikako Tom Boonen buru izango da .
SMatxinT	Flandesko itzulia igande honetan jokatuko, eta beraz, belgikako da Tom Boonen buru izango da.
Oracle	Flandesko itzulia igande honetan eztabaidatzen da, eta beraz, Tom Boonen belgikarra taldeko liderra izango da.

Current results

- Results with those simple features are close to individual systems.
- Oracles show large room for improvement.
- RBMT phrases are underused.
- Current features are not discriminative enough.

Work in progress

- Design of new and more distinctive features for the final decoder.
- Use of multiple trees obtained with different parsers.
- Promote the use of RBMT phrases.
- Use reranking techniques to score the resulting *n*-best lists.

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.

Conclusions

Hybrid translation of patents

The final hybrid translator's elements are being designed and tested independently.

A definite corpus and a concrete domain is needed in order to develop some other components such as the domain grammar and build the systems.

Conclusions

Hybrid translation of patents

The final hybrid translator's elements are being designed and tested independently.

A definite corpus and a concrete domain is needed in order to develop some other components such as the domain grammar and build the systems.

By the way, we need a **name** for our hybrid translator!

SMT Applied to the Patent Domain. Hybridisation with GF and RBMT Paradigms.

Cristina España-Bonet and Lluís Màrquez

Universitat Politècnica de Catalunya, TALP Research Center

– First year project meeting –

Göteborg, March 9th, 2011

Conclusions

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>
```

Conclusions

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>
```

Conclusions

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

Conclusions

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.

Conclusions

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Conclusions

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Conclusions

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–

Conclusions

SMatxinT: Experimental setting

Translation engines

- **RBMT**. Matxin with the Freeling parser
- **SMT**. Moses

Hybrid model

- **SMatxinT**. 1RBMT+1SMT+CD SMT

Conclusions

SMatxinT: Experimental setting, 1RBMT+1SMT+CD SMT

(1)	(2)	(5)	(3)	(4)
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**
Ell	veurà	demà	la jugada	al cine principal

Conclusions

SMatxinT: Experimental setting, 1RBMT+1SMT+CD SMT

(1)	**(2)**	**(5)**	**(3)**	**(4)**
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**

Ell	veurà	demà	l'obra	al teatre principal
ϕ	mirarà		la jugada	al cinema principal
	...		...	al teatre del centre
			l'obra al cinema del centre	
			l'obra al teatre principal	
			...	

...

Anirà a veure demà l'obra al teatre principal
Ell mirarà demà la jugada al teatre principal
...

Conclusions

SMatxinT: Experimental setting, 1RBMT+1SMT+CD SMT

(1)	**(2)**	**(5)**	**(3)**	**(4)**
He	**is going to see**	**tomorrow**	**the play**	**in the main theatre**

Ell	veurà	demà	l'obra	al teatre principal
			l'obra al cinema del centre	

...

Anirà a veure demà l'obra al teatre principal

Conclusions

SMatxinT: Experimental setting

Features (for the final monotonous decoder)

Standard features

- Language model
- Word penalty
- Phrase penalty

Binary system features

- SMT (1/0)
- RBMT (1/0)
- Both (1/0)

Conclusions

SMatxinT: Experimental setting

Corpora (Spanish-to-Basque)

- **SMT Training**. Administrative domain (8 Mwords)
- **SMT Development**. Administrative domain (1500 sentences)
- **Test ADMIN**. Administrative domain (1000 sentences)
- **Test EITB**. News domain (1000 sentences)

Conclusions

SMatxinT: Output analysis

Example 1

SPA	Además de mostrar su indignación, los concentrados exigieron el fin de este tipo de violencia.
EUS	Haserrea erakusteaz gain, kontzentrazioan parte hartu zutenek indarkeria mota honen bukaera eskatu zuten.
RBMT	Haren haserrea erakutsi gain, kontzentratuek indarkeriaren mota honen bukaera eskatu zuten.
SMT	Beren gain, eskatu kontzentratuak haserretu da, horrelako indarkeria.
SMatxinT	Beren gain, eskatu kontzentratuak haserretu da, horrelako indarkeria.
Oracle	Bere haserrea erakutsi gain, los kontzentratuak indarkeria mota honen bukaera eskatu zuten .