

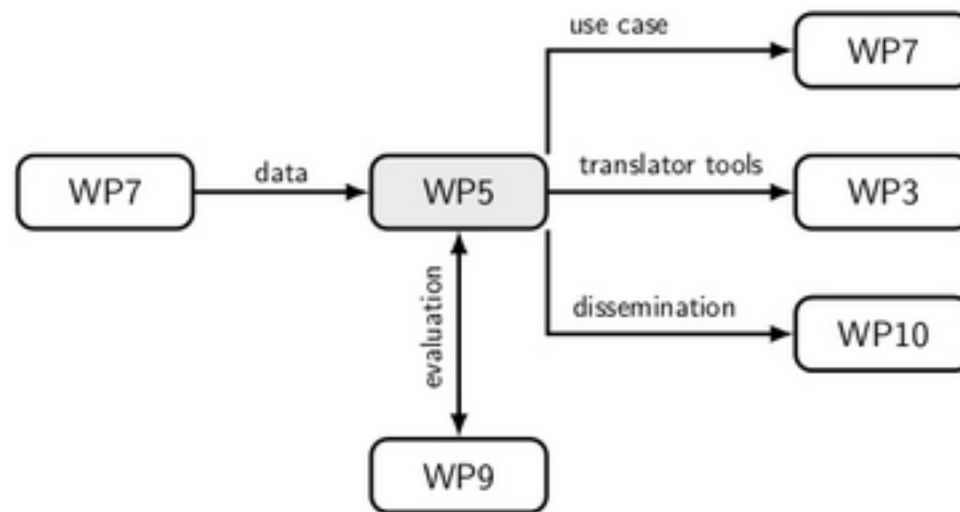
Patent translation

Goals

Explore and build translation engines specialised for patent translation

Integrate the translations into the patents retrieval system

Division of work



Patents

Patent documents

Meta-information

IPC classification A61P

Specific therapeutic activity of chemical compounds or medical preparations.

```
-<patent-document uid="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" uid="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" uid="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-ipc>
<classification-ipc mxw-id="PCL624787575" load-source="docdb" status="new">A61K 31/135 20060101C I20051008RMEP </classification-ipc>
<classification-ipc mxw-id="PCL624787849" load-source="docdb" status="new">A61P 3/04 20060101ALI20051220RMJP </classification-ipc>
<classification-ipc mxw-id="PCL624795950" load-source="docdb" status="new">A61K 31/135 20060101A I20051008RMEP </classification-ipc>
<classification-ipc mxw-id="PCL624799973" load-source="docdb" status="new">A61P 25/20 20060101ALI20051220RMJP </classification-ipc>
<classification-ipc mxw-id="PCL624806558" load-source="docdb" status="new">A61K 31/137 20060101CFI20071018BHEP </classification-ipc>
<classification-ipc mxw-id="PCL624810330" load-source="docdb" status="new">A61K 31/137 20060101AFI20071018BHEP </classification-ipc>
<classification-ipc mxw-id="PCL624820189" load-source="docdb" status="new">A61P 3/00 20060101CLI20051220RMJP </classification-ipc>
<classification-ipc mxw-id="PCL624827390" load-source="docdb" status="new">A61P 25/00 20060101ALI20071018BHEP </classification-ipc>
<classification-ipc mxw-id="PCL624828540" load-source="docdb" status="new">A61P 25/00 20060101CFI20071018BHEP </classification-ipc>
```

Patent documents

Text

Abstracts and 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 b
  Overall, the mean decrease in body weight for the venlafaxine group at week 10 was 7.5 lb with a mean percent decrease from basel
  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
  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>
```

Patent documents

Language

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

- 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** .

Corpus

Table here

Pre-process

Different process for the translation engine and retrieval system

- Common step: tokenising
- Main difference: mark-up and semantic annotations

Tokenisation

Esquema + Exemple

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

vs.

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

Translation engines

Engines

Plot with SMT, GF => HYBRID

SMT for biomedical patents

Standard SMT system with

- **Corpus:** pre-processed 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

Evaluation in the biomedical domain

Syntactic metrics for MT evaluation

- *MALT* dependency parser for English and French
- *Berkeley* parser for German
- Similarity is computed as the overlap of the linguistic elements in the reference and the candidate.
- Linguistic elements can be either the lexical items, or the results of the parse, such as part-of-speech and phrase constituents.

SMT, automatic evaluation

En2Fr & En2De results

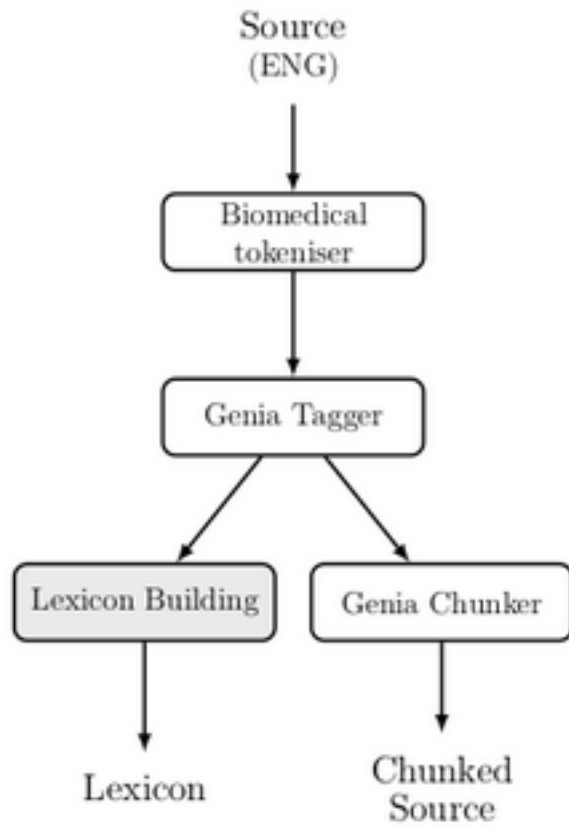
Also other language pairs?

GF for biomedical patents

- Lexicon
- Grammar

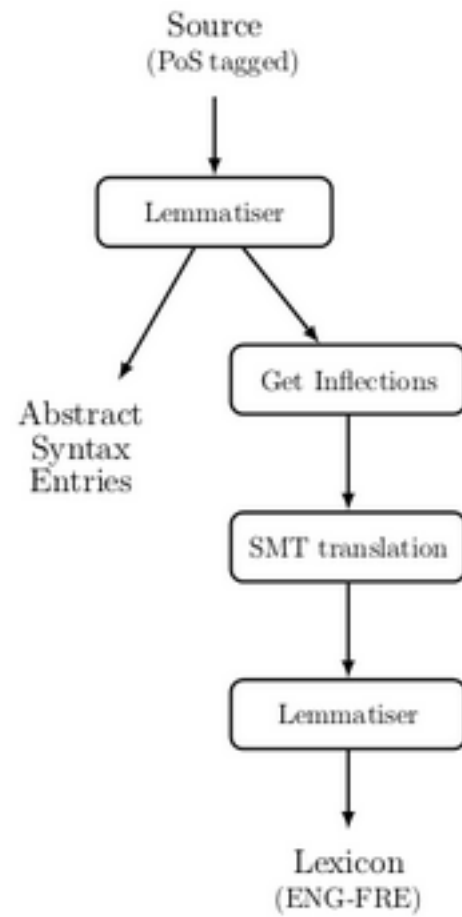
Translation by chunking

Methodology



Lexicon building

Methodology



Lexicon building

German lexicon

***nucleotide sequence* -> Nucleotidsequenz**

Word-to-word GIZA alignments not enough

Solution adopted:

Split compounds, word-to-word mapping, join afterwards

Lexicon building

Static vs. Runtime lexicons

RAMONA something here, please

Construction?

Lexicons for French and German

Sizes and sources for static, safe, unsafe, parse, noparse

RAMONA, please

French concrete grammar

Specific issues

- NPs and AdvP are mapped into GF categories and linearised
- VP, RelP and AdjP are linked to a NP in order to be linearised
- Disambiguation of multiple linearisations by frequency counts in the corpus

French concrete grammar

Table with % of chunks translated

I need to choose only the representative systems (3?)

German concrete grammar

Specific issues

Nominalisation

immunising the mouse-> das Immunisieren von der Mouse

Gerund translated into infinitive + preposition (+ article)

Relative sentences

Pharmaceutical composition comprising an aqueous solution

Gerund and participle sentences not common in German

They are replaced by a relative clause during chunking

German concrete grammar

Table with % of chunks translated

As before I need to choose only the representative systems (3?)

GF, automatic evaluation

Evaluation with lexical and syntactic metrics

1008 fragments from the MAREC test set

...

GF, automatic evaluation for En2Fr

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
GF-StBs	67.02	57.78	65.74	19.91	4.74	18.24	30.78	21.18	56.20
GF-StEs	67.02	57.78	65.74	19.91	4.74	18.24	30.78	21.18	56.20
GF-SaBs	67.96	52.27	65.76	20.41	4.98	18.09	33.64	23.30	60.02
GF-SaEs	66.62	51.14	64.47	21.69	5.17	18.87	35.36	26.20	64.45
GF-UnBs	68.02	52.42	65.84	20.31	4.97	18.04	33.54	23.19	59.75
GF-UnEs	66.68	51.30	64.55	21.57	5.15	18.78	35.25	26.06	64.13

	CP-Oc(*)	CP-Op(*)	CP-STM-9	SP-Op(*)	SP-pNIST-5	ULC
GF-StBs	21.55	22.00	17.98	21.65	1.66	84.11
GF-StEs	21.55	22.00	17.98	21.65	1.66	84.11
GF-SaBs	25.13	25.95	20.58	24.44	1.91	97.07
GF-SaEs	25.92	26.83	21.06	25.21	1.97	99.98
GF-UnBs	25.14	25.82	20.63	24.35	1.89	96.70
GF-UnEs	25.89	26.69	21.09	25.12	1.94	99.55

GF, automatic evaluation for En2De

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
GF-StBs	84.21	76.66	83.52	15.07	3.44	14.26	23.23	12.44	44.35
GF-StEs	84.21	76.66	83.52	15.07	3.44	14.26	23.23	12.44	44.35
GF-SaBs	73.69	62.52	72.25	20.74	4.72	18.49	32.69	20.42	68.05
GF-SaEs	75.58	64.75	74.49	19.69	4.50	17.75	31.23	18.75	63.78
GF-UnBs	74.21	63.04	72.77	20.39	4.67	18.24	32.26	20.12	66.97
GF-UnEs	76.00	65.12	74.91	19.48	4.47	17.61	30.94	18.54	63.06

	CP-Oc(*)	CP-Op(*)	CP-STM-9	SP-Op(*)	SP-pNIST-5	ULC
GF-StBs	13.41	12.91	8.42	12.91	2.82	69.03
GF-StEs	13.41	12.91	8.42	12.91	2.82	69.03
GF-SaBs	20.02	19.04	12.15	19.04	3.82	99.87
GF-SaEs	19.02	18.52	12.13	18.52	3.85	97.89
GF-UnBs	19.77	18.81	12.04	18.81	3.79	98.75
GF-UnEs	18.83	18.40	12.10	18.40	3.83	97.30

GF, robust parsing with patents

Robust parsing applied to patents

Pre-process and cleaning

From

The use of claim 23 , wherein the amount of said composition is from 100 mg to 800 mg of ibuprofen .

To

the use of claim 2 3 wherein the amount of said composition is from 1 0 0 mg to 8 0 0 mg of ibuprofen

Pre-process necessary for parsing

GF, robust parsing with patents

Parsing

- With C Runtime, parseEng, DictEng, ExtraLex
- *Advantages*: robustness
- *Disadvantages*: cleaning and length (<26 tokens)

Linearisation

- With parseGer, DictGer, ExtraLexGer

Use of *generic resources* (parseEng, DictEng, parseGer, DictGer) and *domain lexicons* (ExtraLex, ExtraLexGer)

GF, robust parsing evaluation

Experiment

Marec test set, 1008 fragments

- Cleaning: 537 fragments
- Properly linearised: 98 fragments
- Evaluation with lexical and syntactic metrics

GF, robust parsing evaluation

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
GF-Robust	82.09	64.70	81.25	7.69	2.51	19.76	23.71	16.21	42.33
GF-StBs	81.12	74.43	80.60	10.69	2.59	18.77	23.01	9.00	38.05
GF-StEs	81.12	74.43	80.60	10.69	2.59	18.77	23.01	9.00	38.05
GF-SaBs	70.34	61.00	69.24	19.63	3.78	26.35	35.55	20.73	68.39
GF-SaEs	72.23	62.75	71.84	17.02	3.54	24.72	33.33	19.06	62.39
GF-UnBs	70.47	61.19	69.37	19.33	3.76	26.14	35.32	20.60	67.79
GF-UnEs	72.16	62.75	71.77	17.17	3.55	24.82	33.41	19.06	62.61

	CP-Oc(*)	CP-Op(*)	CP-STM-9	SP-Op(*)	SP-pNIST-5	ULC
GF-Robust	19.32	14.20	12.75	14.20	2.49	73.22
GF-StBs	15.03	12.76	10.08	12.76	2.35	62.82
GF-StEs	15.03	12.76	10.08	12.76	2.35	62.82
GF-SaBs	24.00	21.07	15.76	21.07	3.40	99.19
GF-SaEs	23.13	20.98	16.16	20.98	3.44	99.01
GF-UnBs	23.85	20.97	15.66	20.97	3.37	98.56
GF-UnEs	23.14	21.01	16.25	21.01	3.44	99.17

Further hybridisation

SMT & GF integration lead by *GF*

- GF grammar with SMT built lexicon and disambiguation by frequency counts
- Robust parsing with statistical models for searching the space and for disambiguation

Further hybridisation

SMT & GF integration lead by *SMT*

Additional SMT decoding on top of GF and SMT to choose the best translation options

- **Hard Integration** -- GF phrases are forced to appear -- SMT complements -- top SMT reorders
- **Soft Integration** -- GF and SMT phrases interact -- top SMT reorders and chooses the best option -- LM plays an important role in choosing

Further hybridisation

SMT & GF integration lead by *SMT*

- **Integration only at decoding time** Either Soft or Hard, it is applied on the test set
- **MERT with GF** The final decoder weights are obtained also with an integration in development

Hybrid system

Final system

Characteristics and options

- static vs. dynamic lexicon (two types)
- base vs. extended lexicons
- single vs. multiple GF translations available
- hard vs. soft integration
- integration at decoding time vs. tuning

Hybrid system

Number of phrases from every system choosen at the end

	GF	SMT	BOTH	Total
HIddev-StBs	1,486 (34.92%)	0 (0.00%)	2,769 (65.08%)	4,255
HIddev-SaBs	3,228 (52.65%)	1 (0.02%)	2,902 (47.33%)	6,131
HIddev-StEs	1,486 (34.92%)	0 (0.00%)	2,769 (65.08%)	4,255
HIddev-SaEs	3,242 (50.57%)	1 (0.02%)	3,168 (49.42%)	6,411
HIddev-StEm	1,435 (33.73%)	0 (0.00%)	2,820 (66.27%)	4,255
HIddev-SaEm	2,683 (41.85%)	1 (0.02%)	3,727 (58.13%)	6,411
SIddev-StBs	250 (5.88%)	1,897 (44.58%)	2,108 (49.54%)	4,255
SIddev-SaBs	323 (5.27%)	3,656 (59.63%)	2,152 (35.10%)	6,131
SIddev-StEs	251 (5.90%)	1,906 (44.79%)	2,098 (49.31%)	4,255
SIddev-SaEs	354 (5.52%)	3,737 (58.29%)	2,320 (36.19%)	6,411
SIddev-StEm	230 (5.41%)	1,936 (45.50%)	2,089 (49.10%)	4,255
SIddev-SaEm	438 (6.83%)	3,269 (50.99%)	2,704 (42.18%)	6,411

Hybrid system

Automatic evaluation En2Fr

Table with the best systems

Hybrid system

Automatic evaluation En2De

Table with the best systems

Manual evaluation

Manual evaluation

Setup

Experiment definition

JUSSI, after the evaluation

Manual evaluation

Results

Table?

JUSSI, after the evaluation

Manual evaluation

Conclusions

JUSSI, after the evaluation

Patent translator usage

Patent translator usage

- One-click system
- Offline translation in the retrieval system
- Translation tools?
- Webservice?

One-click system

Perl script that runs the translator

```
csmisc14:hybrid cristina$ perl H1PTrad.pl
```

```
Usage: perl H1PTrad.pl -v # -m [runtime|unsafe|demo] <input> [src2trg]
```

```
-v: verbosity [0,1,2]
```

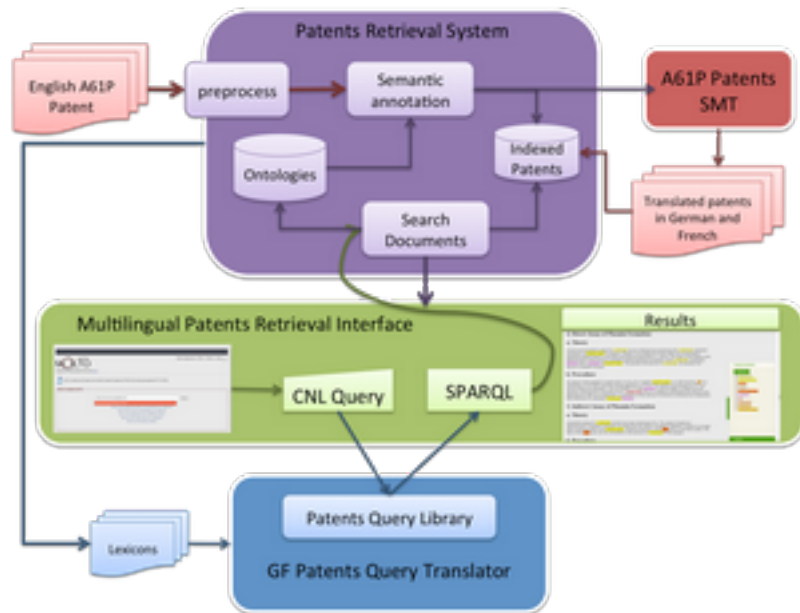
```
-m: mode [runtime|unsafe|demo]
```

```
input: file to translate
```

```
src2trg: language pair
```

```
Ex: perl H1PTrad.pl -v 1 -m demo /Users/systems/input/patsA61P.test.en en2fr
```

Patent translation & retrieval Architecture



- SMT-based pipeline for automatic translation of annotated documents.
- multilingual document retrieval, discussed in the query flagship.
- GF-based querying subsystem for automatic translation of CNL queries to SPARQL. Further discussed in the query flagship.
- User Interface, shown as the case study in the query flagship.

Patent translation & retrieval

Dataset

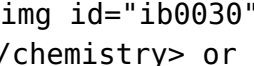
- 7,705 documents, dated 2010 to 2012, downloaded from the EPO website
- 4,485 of them have claims, description and/or abstracts in English, the selected language to annotate the documents

Documents Claims Descriptions Abstracts

English	4,485	62,638	3,832	2,518
German	2,047	32,007	192	80
French	2,011	31,487	130	44

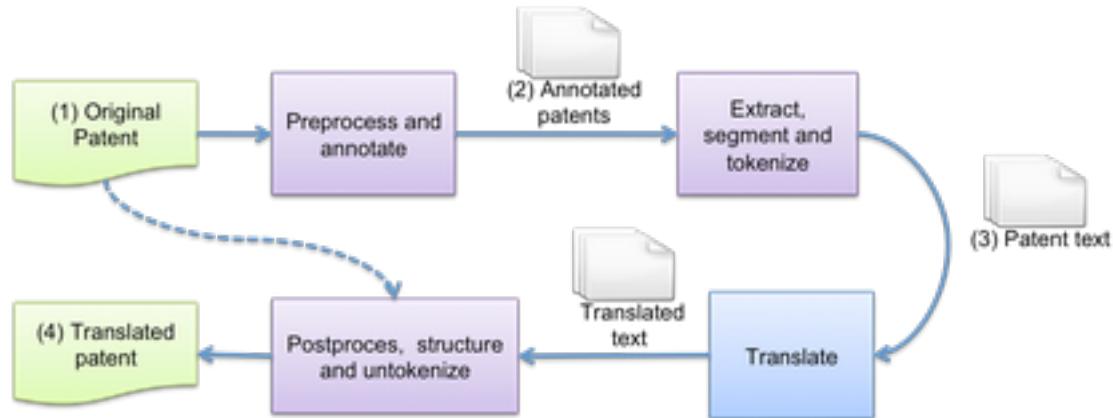
Patent translation & retrieval

Semantic annotations and UTF-8 encode

The use of a compound of the formula:  or isomers i.e. geometric, optical, entianomeric, diasteriomeric, epimeric, stereoisomeric, tautomeric, conformational, or anomeric forms, salts, solvates and chemically protected forms thereof, in the preparation of a medicament for inhibiting the activity of `<AnatomicalStructure inst="umls/id/C1538577" class="semanticnetwork/id/T017">PARP` `</AnatomicalStructure>` , wherein: `<claim-text>`A and B together represent a fused aromatic ring, optionally substituted with one or more substituent groups selected from halo, nitro, hydroxy, ether, thiol, thioether, amino, C $_{1-7}$ alkyl, C $_{3-20}$ heterocyclyl and C $_{5-20}$ aryl; `</claim-text>` `<claim-text>`R C is -CH $_2$ -R L , where R L is a C $_{5-20}$ aryl group, optionally substituted with one or more substituent groups selected from C $_{1-7}$ alkyl, C $_{5-20}$ aryl, C $_{3-20}$ heterocyclyl, halo, hydroxy, ether, nitro, cyano, acyl, carboxy, ester, amido, amino, sulfonamido, acylamido, ureido, acyloxy, thiol, thioether, sulfoxide and sulfone; and `</claim-text>` `<claim-text>`R N is hydrogen. `</claim-text>`

Patent translation & retrieval

Offline translation of the full dataset



- Cleaning and markup
- text extraction, tokenization and segmentation
- translation - Retraining of a new SMT using UTF-8 encoding
- postprocess, XML formatting and merge (EN,DE,FR)

Patent translation & retrieval

Online API

- <http://falkor.lsi.upc.edu/MOLTO/>
- Allows to upload a single file. It should contain text in English and annotations.
- It returns the same document with the english sections translated into German and French.

Patent translation & retrieval

The patents retrieval prototype

- <http://molto-patents.ontotext.com>
- The interface is available in EN, DE and FR

Patent translation & retrieval

English text

The screenshot shows a web browser displaying a patent document from multa-patents.onotext.com/document/EP2068873B1. The document text is in English and describes a pharmaceutical composition. It includes a search bar at the top right with the text "selenium sulfide" and a "1 of 1" indicator. A sidebar on the right titled "Semantic Annotations" contains a list of terms with checkboxes: "Uncheck all", "activesingredient", "anatomicalstructure", "applicant", "applicationdate", "applicationnumber", and "diseaseordysfunction". The main text contains several terms highlighted in green, including "penicillin G", "oxacillin", "ampicillin", "nafcillin", "ticarcillin", "amoxicillin", "cephalosporins", "cephalothin", "cephalexin", "cefazolin", "cephadrine", "cephapirin", "cefamandole", "cefotaxime", "cefoperazone", "cefotaxime", "ceftriaxone", "monobactams", "clavulanic acid", "sulbactam", "tazobactam", "carbapenems", "imipenem", "bactracine", "aminoglycosides", "neomycin", "gentamicin", "clindamycin", "tobramycin", "amikacin", "netilmicin", "lincomycin", "spectinomycin", "erythromycin", "azithromycin", "clarithromycin", "azoles", "metronidazole", "mupirocin", "silver sulfadiazine", "cyclopirox", and "selenium sulfide".

French text

antibactérien agent thérapeutique et d'un agent thérapeutique antifongique

Dans un milieu de réalisation de l'invention, la composition pharmaceutique comprend en outre un ou plusieurs agents antibactériens. Non-limiting exemples de antibacterial, agents sont bêta-lactames, tels que les pénicillines (par exemple, pénicilline G, l'oxacilline, l'ampicilline, la nafcilline, la ticarcilline, et l'amoxicilline), les céphalosporines (par exemple, la céphalothine, la céphalexine, cefazoline, céphadrine, la céphapirine, cefamandole, céfoxidine, cefoperazone, céfotaxime, et de la céftriaxone), monobactams (par exemple, l'aztréonam), la bêta-lactamase inhibiteurs (par exemple, l'acide clavulanique, sulbactam, et tazobactam), les carbapenems et (par exemple, imipénème), ainsi que des dérivés de tels bêta-lactames; polypeptides, tels que bactracine; aminoglycosides, tels que la néomycine, la gentamicine, la clindamycine, tobramycine, l'amikacine, la netilmicine, et lincomycine; amino-glycosidic-de type compounds, tels que spectinomycine; macrolides, tels que l'erythromycine, la streptomycine, l'azithromycine, et de clarithromycine; des azoles tels que metronidazole; de mupirocine; azines, tels que

Webservice?

Something here?

Translator tools

screenshot if integrated

Prova

First Header | Second Header ----- | ----- Content Cell | Content Cell Content Cell |
Content Cell