# Challenges when Translating Scientific Documents

presented by Holger Schwenk

LIUM, University of Le Mans, France

*Holger.Schwenk@lium.univ-lemans.fr*

November 07, 2012

COSMAT

partenaires :
INRIA LIUM ANR SYSTRAN

# Plan

- Translating scientific documents
- COSMAT project
- Statistical machine translation
- Adapting the system to scientific documents
- Outlook



# Scientific Documents and Language

- English is the universal language for scientific communication
- A large part of the researchers are not native speakers of English
- Regulations in many countries require researches to publish in their native language
- Situation in France:
  - PhD thesis, progress or government reports must be written in French
  - French is the official educational language

$\Rightarrow$ Researches frequently translate between English and their own language



# Use Cases

## French Scientist

- Often publishes similar texts in French and English
- His translation efforts could be leveraged

## International researcher

- Does not consider publications in French (PhD thesis, etc)
- Important parts of national research activities are not visible at the international level

## French public

- Interested in (English) publications
- But standard on-line translation tools are not appropriate



# Machine Translation of Scientific Documents

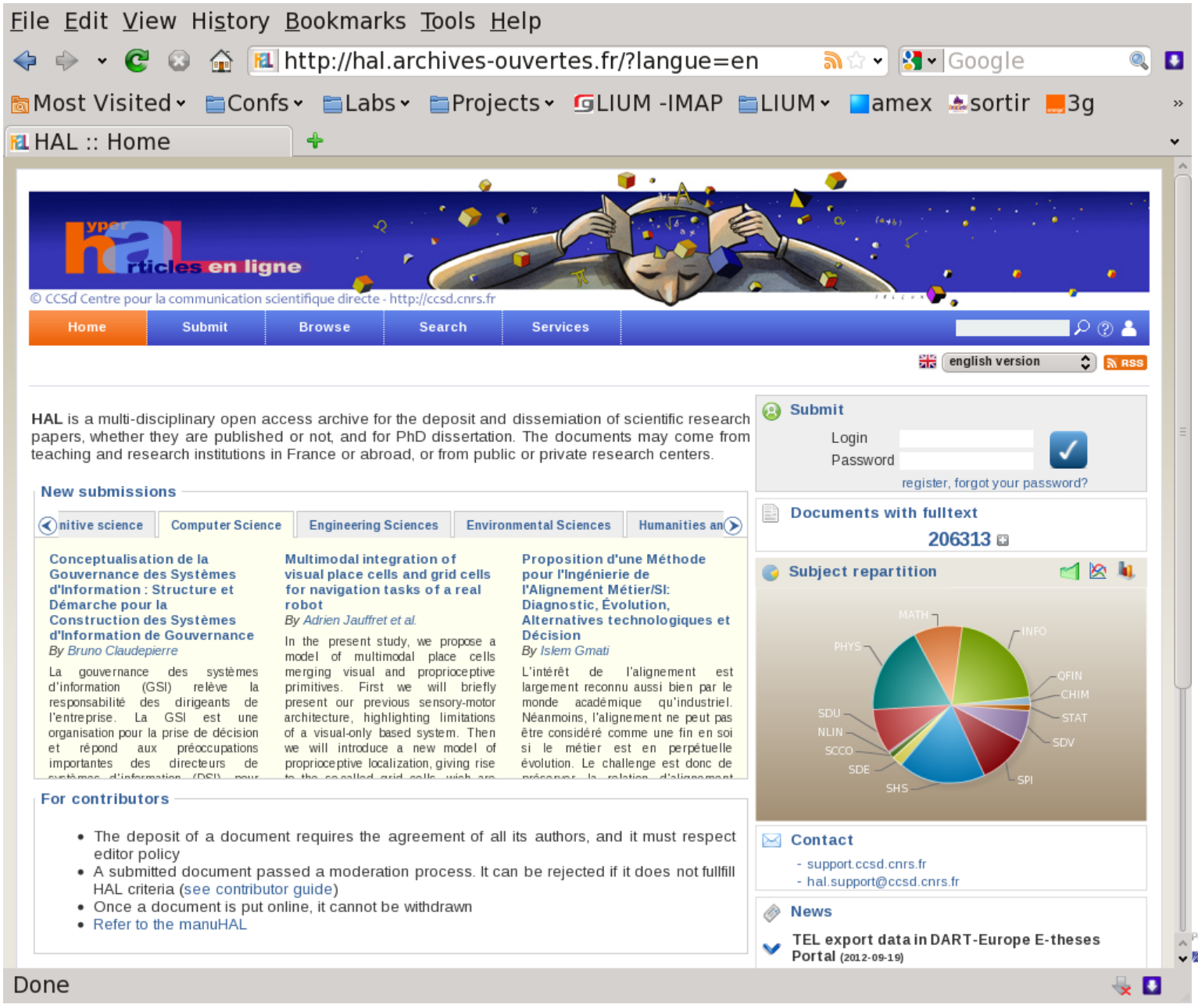
- Each scientific domain has its own terminology
- Example: translation of the English word *cluster*
  - *grappe* (computer science: cluster of machines)
  - *regroupement* (math: clustering of data)
  - *amas* (astronomy: cluster of stars)

$\Rightarrow$ A general purpose MT systems is unlikely to provide appropriate translations

$\Rightarrow$ Need for domain specific models or automatic adaptation



# The HAL Open Archive



# The HAL Open Archive

- Created in 2006
- Multi-disciplinary open access archive for the deposit and dissemination
  - of scientific research papers (published or not)
  - PhD thesis (TEL)
- Documents may come from teaching and research institutions in France or abroad
- From public or private research centers
- About 30 technical domains (labeled by the authors)
  - Computer Science (21%)
  - Physics and Social Sciences (18% each)
  - Math and engineering Sciences (10% each)
- Currently 206k documents in PDF



# The French COSMAT Project

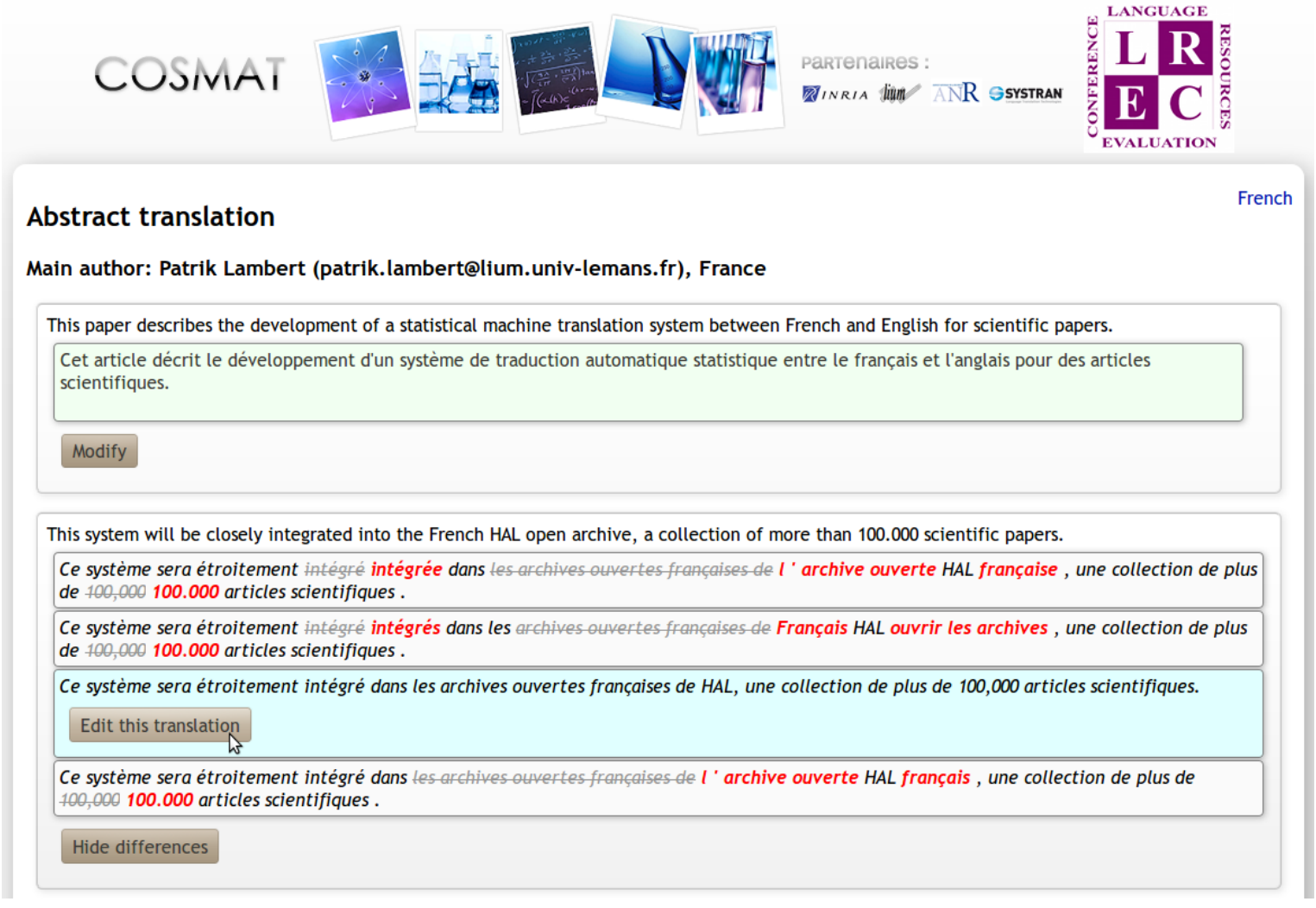
## Objectives

- Smoothly integrate MT into the HAL work flow
- Analyse PDF document and automatically extract meta-information like title, authors, etc
- Put those values into the HAL forms
- Provide an automatic translation of the document (statistical and hybrid)
- Interactive interface to post-edit the provided translation
- Incremental update of the MT systems

## Characteristics

- Duration: 10/2009 – 3/2013
- Partners: LIUM, SYSTRAN and INRIA

# Interactive User Interface



# History of Machine Translation

- Machine translation is one of the oldest research areas in computer science

Sentences in Russian are punched into standard cards for feeding into the electronic data processing machine for translation into English

- First system by IBM in 1954 (Georgetown) : translation of 60 sentences between Russian and English

$\Rightarrow$ Great euphoria and many research projects

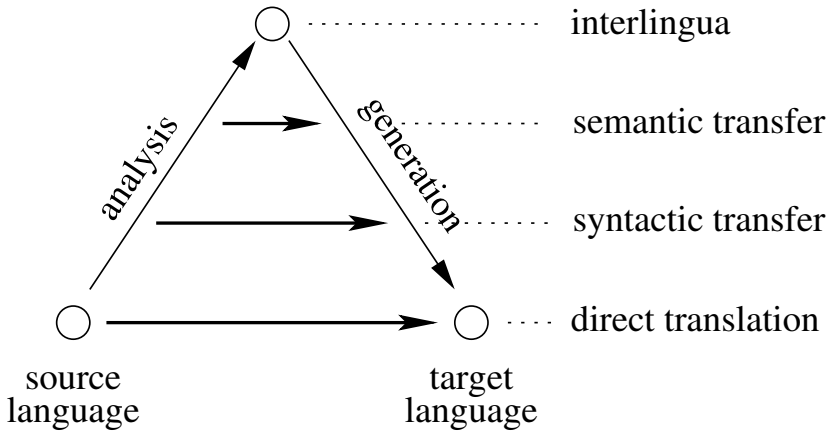
- Results did not reach the expectations ( « ALPAC report » in 1966 )



# History of Machine Translation

## First approaches

- The triangle of Vauquois (1968)




- The techniques are based on an analysis of the sentence, the transfer and the generation of the target sentence
- Important implication of bilingual humans
- Quite difficult to deal with a large amounts of languages


# Statistical Approaches in MT ?

- MT deals with words
- There are grammatical relations between the words, they have different meanings, etc
- How to apply statistical approaches to MT ?
- We need to make decisions:
  - choice of the words
  - use of particular expressions
  - which word order ?
  - ...
- Usually, there are several choices, ambiguities, etc

$\Rightarrow$ This is best taken care of with statistics and probabilities !



# The Statistical Approach to MT

Traduction par approche statistique = Modélisation linguistique + Prise de décisions par des méthodes statistiques

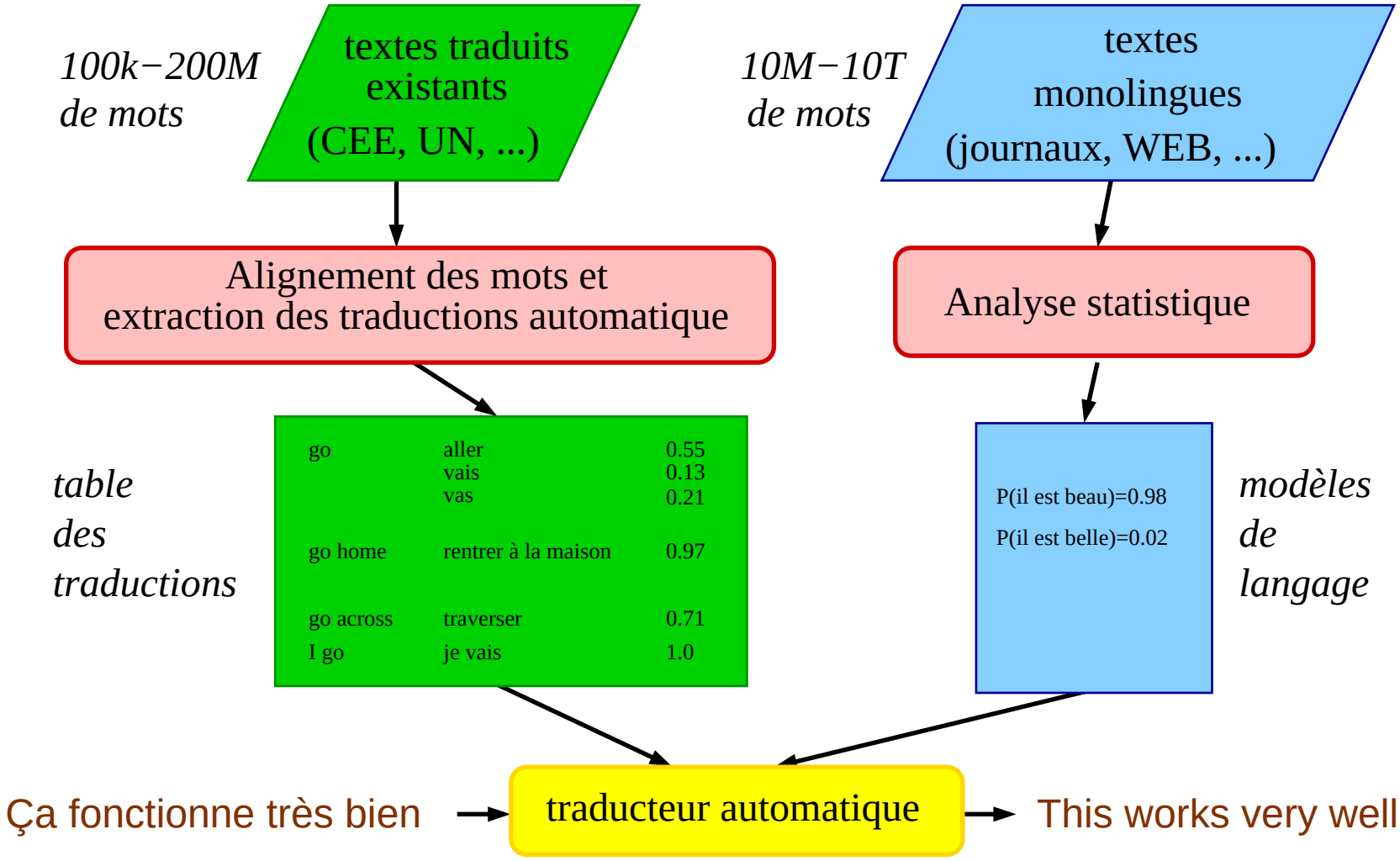
- Translation of a French sentence $f$ into English $e$

$$e^* = \arg\max_e P(e|f) = \arg\max_e P(f|e)P(e)$$

- Statistical translation model
- Statistical language model
- Statistical decision



# The Statistical Approach to MT



# Monolingual Corpora

## Potential sources

- All text in the target language
- Usually, it's easy to find monolingual texts
- Even in special domains
- We dispose of many years of news paper texts

$\Rightarrow$ several billions of words (30 km of dictionaries)

- The WEB is an almost unlimited resources of texts in many languages



# Parallel Data

## Potential sources

- International organisations : EC, UN, ...
- Often written in special language or jargon
- Multilingual Internet site: Wikipedia, press agencies, manuals, ...

## Specific sources

- Bilingual domain specific texts are a very useful resource
- Unfortunately, they are rarely (freely) available for research

$\Rightarrow$ The amount, quality and type of resources heavily influence the performance on an SMT system



# Domain Specific Resources

- HAL contains PhD thesis with bilingual abstracts entered by the researchers
- Extracted, cleaned and aligned to create parallel corpora
  - Procedure limited to the domain of computer science and physics
  - Development and test: 1000 sentences, 30k words
  - Training: 56k sentences, 1.5M words
- Extraction of monolingual data from PDF documents
  - Improvement of Grobid tool to convert text to TEI format
  - English: 104M, French: 36M words
- This data is freely available to foster research on translation of scientific documents and domain adaptation



# Adapting a Generic SMT System

## Adaptation

- Data selection
- Data weighting
- Better probability estimation
- Unsupervised training

## User feedback

- Fast incremental update



# Data Selection

## Principle

- Extract only the relevant data from all the available one
- This needs a criterion on the relevance of the data

## Approach

- Train a language model on a in-domain and generic corpus
- Use these language models to judge each sentence in the corpus
- Sort the sentences according to this criterion

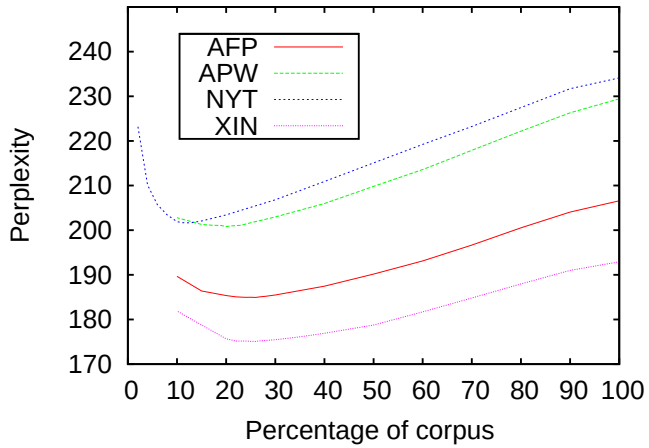
$\Rightarrow$ Keep only the most relevant ones



# Data Selection in Practice

## Monolingual Data

- Data selection is performed for each corpus
- Perplexity reduction of up to 15%




- Between 14 and 26% of the data is selected

$\Rightarrow$ Smaller and better models

## Parallel Data

- Same procedure on source or target only
- Extension to bilingual data [Axelrood et al, EMNLP'11]


# Bitext Weighting

- Data selection is an effective method, but it is a binary decision
- Parallel data is heterogeneous w/r to size, appropriateness, quality, ...
- Weighting could be performed at various levels:
  corpus, sentence, phrase
- Weighting LM training data (at the corpus level) is a well established procedure
- There seems to be no best practice for the translation model
- Generic approach aiming to fix the probability distributions instead of adding various features



# Continuous Space Methods

## Probability models in SMT systems

- Current language and translation models list all extracted knowledge
- Probability estimation is performed in a discrete space
- Generalisation to unobserved events is not straight-forward

## Alternative approach

- Project word indices onto a continuous space and use a probability estimator operating on this space
- Probability functions are smooth functions and better generalization can be expected



# Unsupervised Training

## Motivation

- Bilingual topic specific data is difficult to obtain and limited in size (Cosmat: 1.5M words)
- Monolingual data is usually easier to collect (Cosmat: 100M words)

$\Rightarrow$ use this monolingual data to **adapt** the translation model

## Procedure

- Translate the monolingual data with an MT system
- Filter the translations and keep only the most reliable ones
- Inject this data as additional parallel training data

$\Rightarrow$ The systems synthesis itself new data !

- This procedure can't add new translations, but probability distributions of existing data are shifted towards the topic

# Some Results

## Automatic Evaluation : BLEU score

- Comparison with reference translation
- Precision of short sequences of words
- The higher the BLEU score the better the translation

## Results comparison

| | Computer Science | Physics |
|---|---|---|
| Baseline: | 27.3 | 27.1 |
| Adapted system: | 38.8 | 40.0 |

$\Rightarrow$ very significant improvements



# Example Translations

SRC: To obtain an equivalence accounting for the **multiple interferences**, a generalization of **T-coloring problems for the hypergraphs** is introduced.

BASE: Pour obtenir une équivalence avec les **multiples ingérences**, une généralisation des problèmes pour le **hypergraphs T-coloring**.

ADAPT: Pour obtenir une équivalence représentant les **interférences multiples**, une généralisation du problème de **T-coloration pour les hypergraphes** est introduit.

REF: Pour obtenir une équivalence rendant compte des **interférences multiples**, une généralisation du problème de **T-coloration pour les hypergraphes** est introduite.



# Example Translations

SRC: In our thesis we propose a **test methodology based on** model checking and static analysis.

BASE: Dans notre thèse que nous proposons une **méthodologie pour tester** le modèle **fondé** sur le contrôle et l'analyse statique.

ADAPT: Dans notre thèse nous proposons une **méthodologie de test basée sur** le model checking et de l'analyse statique.

REF: Dans notre thèse on propose une **méthodologie de test basée sur** des techniques issues des domaines de la vérification et de l'analyse statique.


# Example Translations

SRC: The verification of the **cryptographic protocols** ensures that there is not possible **attack** during an execution of the protocol ..

BASE: La vérification des **protocoles de chiffrement** assure qu'il n'est pas possible d'**attentat** lors d'une exécution du protocole ...

ADAPT: La vérification des **protocoles cryptographiques** assure qu'il n'est pas **attaque** possible durant une exécution du protocole ...

REF: La vérification des **protocoles cryptographiques** assure qu'il n'existe pas d'**attaque** possible lors d'une exécution du protocole ...



# Example Translations

SRC: A **Newmark time scheme** is used, together with a finite element mesh and a **domain decomposition method**.

BASE: Un **temps Newmark système** est utilisé, avec un maillage d'éléments finis et une **méthode de décomposition**.

ADAPT: Un **schéma en temps Newmark** est utilisé, avec un maillage d'éléments finis et une **méthode de décomposition de domaine**.

REF: Elle utilise un **schema de type Newmark en temps**, et une discretisation spatiale en elements finis avec **decomposition de domaine**.



# Conclusion

## Conclusion

- Domain adaptation is very effective to achieve good MT of scientific texts
- It is very important to correctly use the available data
- Correct processing and formatting of scientific documents is tricky (equations, formulaes, tables, special characters, etc)
- SYTRAN has developed a pipeline to take care of this



# Conclusion

## Cosmat corpus

- The data extracted form the HAL archive seems to be a valuable resource to develop MT systems for scientific texts
- This corpus is freely available
- In principle, the same procedure could be applied for the other scientific domains of HAL
- It may be also interesting to use the PDF processing tools to process the large collection of papers in the ACL anythology (21.8k papers)



# Conclusion

## We are ready to go on-line with the Cosmat systems

- Valuable feedback on usefulness of MT of scientific documents
- We hope to collect significant amounts of user postedited text
- Further improvements of the MT systems are expected
- This data will be also made available to the community



# Perspectives

Realtime multilingual lecture translation and subtitling



# Merci beaucoup !

Thank you very much !

Vielen Dank !

Muchas gracias !

Grazie mille !

Najlepša hvala !

Heel hartelijk bedankt !

Paljon kiitoksia !

Mange tak !

...