NRC·CNRC
Institute for
Information
Technology

Bike: Bilingual Keyphrase Experiments

David Nadeau, Caroline Barrière and **George Foster**
Institute for Information Technology
National Research Council Canada
Gatineau, Quebec, Canada

National Research Council Canada
Conseil national de recherches Canada

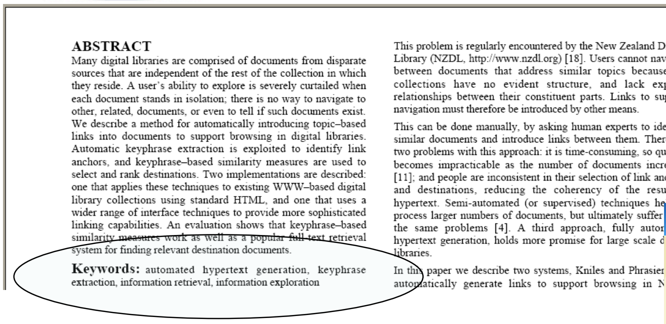
Canada

NRC·CNRC
Institute for
Information
Technology

Introduction

- In this talk, I report experiments we conducted on the task of translating lists of keyphrases.

Translating lists of keyphrases



ABSTRACT

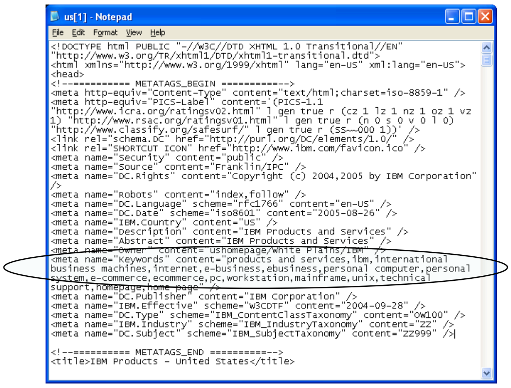
Many digital libraries are comprised of documents from disparate sources that are independent of the rest of the collection in which they reside. A user's ability to explore is severely curtailed when each document stands in isolation; there is no way to navigate to other, related, documents, or even to tell if such documents exist. We describe a method for automatically introducing topic–based links into documents to support browsing in digital libraries. Automatic keyphrase extraction is exploited to identify link anchors, and keyphrase–based similarity measures are used to select and rank destinations. Two implementations are described: one that applies these techniques to existing WWW–based digital library collections using standard HTML, and one that uses a wider range of interface techniques to provide more sophisticated linking capabilities. An evaluation shows that keyphrase–based similarity measures work as well as a popular full-text retrieval system for finding relevant destination documents.

Keywords: automated hypertext generation, keyphrase extraction, information retrieval, information exploration

This problem is regularly encountered by the New Zealand Digital Library (NZDL, http://www.nzdl.org) [18]. Users cannot navigate between documents that address similar topics because the collections have no evident structure, and lack explicit relationships between their constituent parts. Links to support navigation must therefore be introduced by other means.

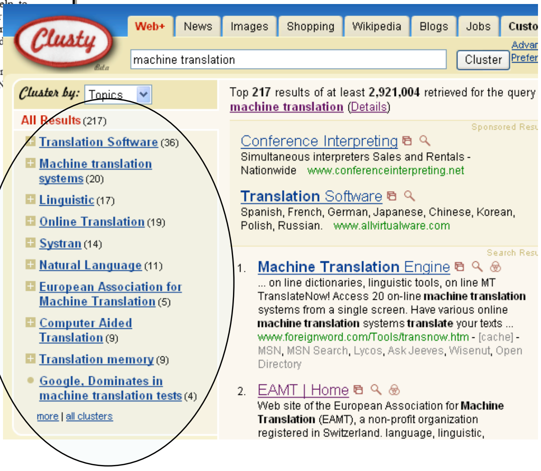
This can be done manually, by asking human experts to identify similar documents and introduce links between them. There are two problems with this approach: it is time-consuming, so quickly becomes impracticable as the number of documents increases [11]; and people are inconsistent in their selection of link anchors and destinations, reducing the coherency of the resulting hypertext. Semi-automated (or supervised) techniques help to process larger numbers of documents, but ultimately suffer the same problems [4]. A third approach, fully auton hypertext generation, holds more promise for large scale d libraries.

In this paper we describe two systems, Kniles and Phrasier automatically generate links to support browsing in N



```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="en-US" xml:lang="en-US">
<head>
<!--========== METATAGS_BEGIN ==========-->
<meta http-equiv="Content-Type" content="text/html;charset=iso-8859-1" />
<meta http-equiv="PICS-Label" content='(PICS-1.1
"http://www.icra.org/ratingsv02.html" l gen true r (cz 1 lz 1 nz 1 oz 1 vz
1) "http://www.rsac.org/ratingsv01.html" l gen true r (n 0 s 0 v 0 l 0)
"http://www.classify.org/safesurf/" l gen true r (SS~~000 1))' />
<link rel="schema.DC" href="http://purl.org/DC/elements/1.0/" />
<link rel="SHORTCUT ICON" href="http://www.ibm.com/favicon.ico" />
<meta name="Security" content="public" />
<meta name="Source" content="Franklin/IPC" />
<meta name="DC.Rights" content="Copyright (c) 2004,2005 by IBM Corporation"
/>
<meta name="Robots" content="index,follow" />
<meta name="DC.Language" scheme="rfc1766" content="en-US" />
<meta name="DC.Date" scheme="iso8601" content="2005-08-26" />
<meta name="IBM.Country" content="US" />
<meta name="Description" content="IBM Products and Services" />
<meta name="Abstract" content="IBM Products and Services" />
<meta name="Owner" content="ushomepage/White Plains/IBM" />
<meta name="Keywords" content="products and services,ibm,international
business machines,internet,e-business,ebusiness,personal computer,personal
system,e-commerce,ecommerce,pc,workstation,mainframe,unix,technical
support,homepage,home page" />
<meta name="DC.Publisher" content="IBM Corporation" />
<meta name="IBM.Effective" scheme="W3CDTF" content="2004-09-28" />
<meta name="DC.Type" scheme="IBM_ContentClassTaxonomy" content="OW100" />
<meta name="IBM.Industry" scheme="IBM_IndustryTaxonomy" content="ZZ" />
<meta name="DC.Subject" scheme="IBM_SubjectTaxonomy" content="ZZ999" />

<!--========== METATAGS_END ==========-->
<title>IBM Products - United States</title>
```



Top **217** results of at least **2,921,004** retrieved for the query **machine translation** (Details)

All Results (217)
- ⊞ Translation Software (36)
- ⊞ Machine translation systems (20)
- ⊞ Linguistic (17)
- ⊞ Online Translation (19)
- ⊞ Systran (14)
- ⊞ Natural Language (11)
- ⊞ European Association for Machine Translation (5)
- ⊞ Computer Aided Translation (9)
- ⊞ Translation memory (9)
- ● Google, Dominates in machine translation tests (4)

more | all clusters

Sponsored Resu

Conference Interpreting
Simultaneous interpreters Sales and Rentals - Nationwide www.conferenceinterpreting.net

Translation Software
Spanish, French, German, Japanese, Chinese, Korean, Polish, Russian. www.allvirtualware.com

Search Resu

1. Machine Translation Engine
... on line dictionaries, linguistic tools, on line MT TranslateNow! Access 20 on-line **machine translation** systems from a single screen. Have various online **machine translation** systems **translate** your texts ...
www.foreignword.com/Tools/transnow.htm - [cache] - MSN, MSN Search, Lycos, Ask Jeeves, Wisenut, Open Directory

2. EAMT | Home
Web site of the European Association for **Machine Translation** (EAMT), a non-profit organization registered in Switzerland. language, linguistic,

Translating lists of keyphrases

- Related to:
  - Sub-sentence-level translation (noun phrases, terminology, …)
  - Query translation (cross-lingual IR);
  - Cross-lingual summarization (Summary as a list of Keyphrases)

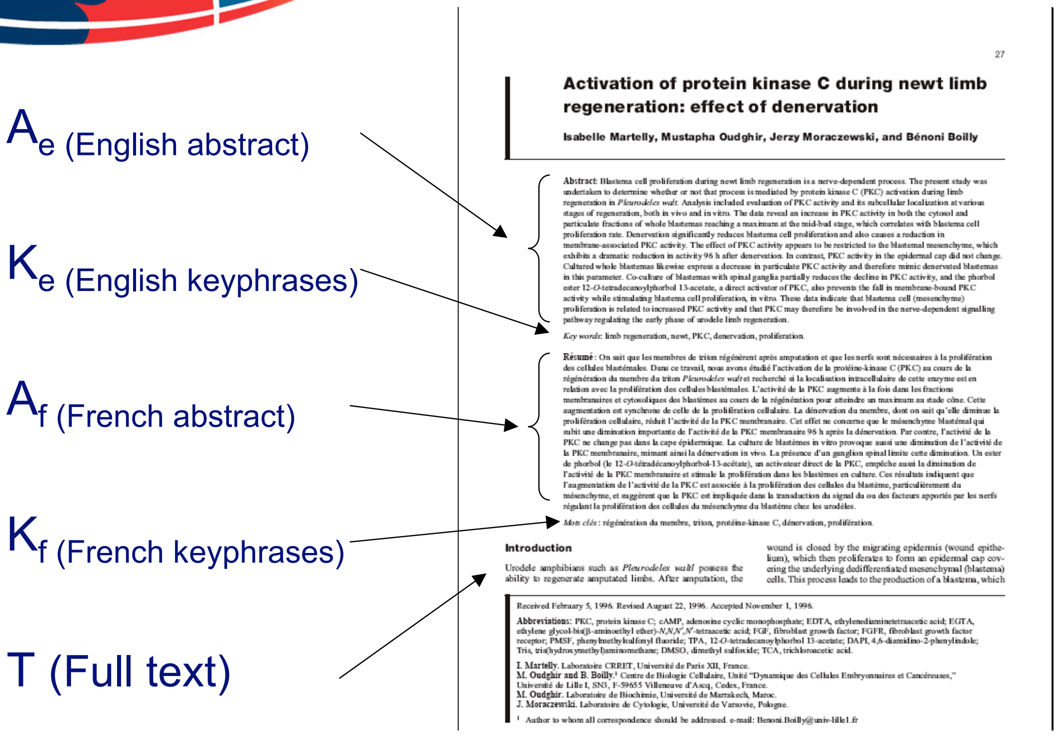
Translating lists of keyphrases

- BIKE (Bilingual Keyphrase Experiments):
  - Bidirectional French ←→ English keyphrase translation
  - Strategy pieces:
    - Statistical MT
    - Terminological resource
    - Inflectional morphology
    - WSD (list of keyphrases as a context)
  - Experiments in training and combining pieces.

Task

- Collection of 3058 scientific papers from the Canada Institute for Scientific and Technical Information (**CISTI**) in ten domains (biochemistry, botany, chemistry, civil engineering, environment (x2), genomics, geotechnical, microbiology and pharmacology)
- Each document is a tuple {A_f, A_e, K_f, K_e, T}

Task



A_e (English abstract)

K_e (English keyphrases)

A_f (French abstract)

K_f (French keyphrases)

T (Full text)

27

Activation of protein kinase C during newt limb regeneration: effect of denervation

Isabelle Martelly, Mustapha Oudghir, Jerzy Moraczewski, and Bénoni Boilly

Abstract: Blastema cell proliferation during newt limb regeneration is a nerve-dependent process. The present study was undertaken to determine whether or not that process is mediated by protein kinase C (PKC) activation during limb regeneration in *Pleurodeles walt*. Analysis included evaluation of PKC activity and its subcellular localization at various stages of regeneration, both in vivo and in vitro. The data reveal an increase in PKC activity in both the cytosol and particulate fractions of whole blastemas reaching a maximum at the mid-bud stage, which correlates with blastema cell proliferation rate. Denervation significantly reduces blastema cell proliferation and also causes a reduction in membrane-associated PKC activity. The effect of PKC activity appears to be restricted to the blastemal mesenchyme, which exhibits a dramatic reduction in activity 96 h after denervation. In contrast, PKC activity in the epidermal cap did not change. Cultured whole blastemas likewise express a decrease in particulate PKC activity and therefore mimic denervated blastemas in this parameter. Co-culture of blastemas with spinal ganglia partially reduces the decline in PKC activity, and the phorbol ester 12-*O*-tetradecanoylphorbol 13-acetate, a direct activator of PKC, also prevents the fall in membrane-bound PKC activity while stimulating blastema cell proliferation, in vitro. These data indicate that blastema cell (mesenchyme) proliferation is related to increased PKC activity and that PKC may therefore be involved in the nerve-dependent signalling pathway regulating the early phase of urodele limb regeneration.

Key words: limb regeneration, newt, PKC, denervation, proliferation.

Résumé : On sait que les membres de triton régénèrent après amputation et que les nerfs sont nécessaires à la prolifération des cellules blastémales. Dans ce travail, nous avons étudié l'activation de la protéine-kinase C (PKC) au cours de la régénération du membre du triton *Pleurodeles walt* et recherché si la localisation intracellulaire de cette enzyme est en relation avec la prolifération des cellules blastémales. L'activité de la PKC augmente à la fois dans les fractions membranaires et cytosoliques des blastèmes au cours de la régénération pour atteindre un maximum au stade cône. Cette augmentation est synchrone de celle de la prolifération cellulaire. La dénervation du membre, dont on sait qu'elle diminue la prolifération cellulaire, réduit l'activité de la PKC membranaire. Cet effet ne concerne que le mésenchyme blastémal qui subit une diminution importante de l'activité de la PKC membranaire 96 h après la dénervation. Par contre, l'activité de la PKC ne change pas dans la cape épidermique. La culture de blastèmes in vitro provoque aussi une diminution de l'activité de la PKC membranaire, mimant ainsi la dénervation in vivo. La présence d'un ganglion spinal limite cette diminution. Un ester de phorbol (le 12-*O*-tétradécanoylphorbol-13-acétate), un activateur direct de la PKC, empêche aussi la diminution de l'activité de la PKC membranaire et stimule la prolifération dans les blastèmes en culture. Ces résultats indiquent que l'augmentation de l'activité de la PKC est associée à la prolifération des cellules du blastème, particulièrement du mésenchyme, et suggèrent que la PKC est impliquée dans la transduction du signal du ou des facteurs apportés par les nerfs régulant la prolifération des cellules du mésenchyme du blastème chez les urodèles.

Mots clés : régénération du membre, triton, protéine-kinase C, dénervation, prolifération.

Introduction

Urodele amphibians such as *Pleurodeles waltl* possess the ability to regenerate amputated limbs. After amputation, the wound is closed by the migrating epidermis (wound epithelium), which then proliferates to form an epidermal cap covering the underlying dedifferentiated mesenchymal (blastema) cells. This process leads to the production of a blastema, which

Received February 5, 1996. Revised August 22, 1996. Accepted November 1, 1996.

Abbreviations: PKC, protein kinase C; cAMP, adenosine cyclic monophosphate; EDTA, ethylenediaminetetraacetic acid; EGTA, ethylene glycol-bis(β-aminoethyl ether)-*N,N,N′,N′*-tetraacetic acid; FGF, fibroblast growth factor; FGFR, fibroblast growth factor receptor; PMSF, phenylmethylsulfonyl fluoride; TPA, 12-*O*-tetradecanoylphorbol 13-acetate; DAPI, 4,6-diamidino-2-phenylindole; Tris, tris(hydroxymethyl)aminomethane; DMSO, dimethyl sulfoxide; TCA, trichloroacetic acid.

I. Martelly. Laboratoire CRRET, Université de Paris XII, France.
M. Oudghir and B. Boilly.[1] Centre de Biologie Cellulaire, Unité "Dynamique des Cellules Embryonnaires et Cancéreuses," Université de Lille I, SN3, F-59655 Villeneuve d'Ascq, Cedex, France.
M. Oudghir. Laboratoire de Biochimie, Université de Marrakech, Maroc.
J. Moraczewski. Laboratoire de Cytologie, Université de Varsovie, Pologne.

[1] Author to whom all correspondence should be addressed. e-mail: Benoni.Boilly@univ-lille1.fr

Task (performance metric)

- English → French
- For each keyphrase, 1 candidate translation is produced
- Exact translation required (reproductibility, lowerbound exp.)
- **Accuracy**: ratio of correctly translated keyphrases to the total number of keyphrases
- Results are reported for all domains (30% held-out split)
- Significance testing done using 10-domain split

Experiment 1

- Baseline (no translation)

Experiment 2

Statistical MT:

- Phrase-based statistical MT [Koehn *et al*., 2003]

- We calculate the conditional Fr-given-En probabilities, and retain only the most probable translation

- Hansard Model: Canadian Hansard parallel corpora
- CISTI Model: 40% training split (abstracts and keyphrases)
  - CISTI « global »: all journals
  - CISTI « individual » one model per journal

Experiment 2

- How to improve statistical MT translation?
- Problem:

English	**French**
fiber	*de* fibre
population sizes	*des* dimensions des populations

- Remove prefixes ("de la ", "le ", "la ", "les ", "l'", "du ", "de ", ",") and suffixes (" ," , " de", " du", “ des") from French translation proposed by the MT system.

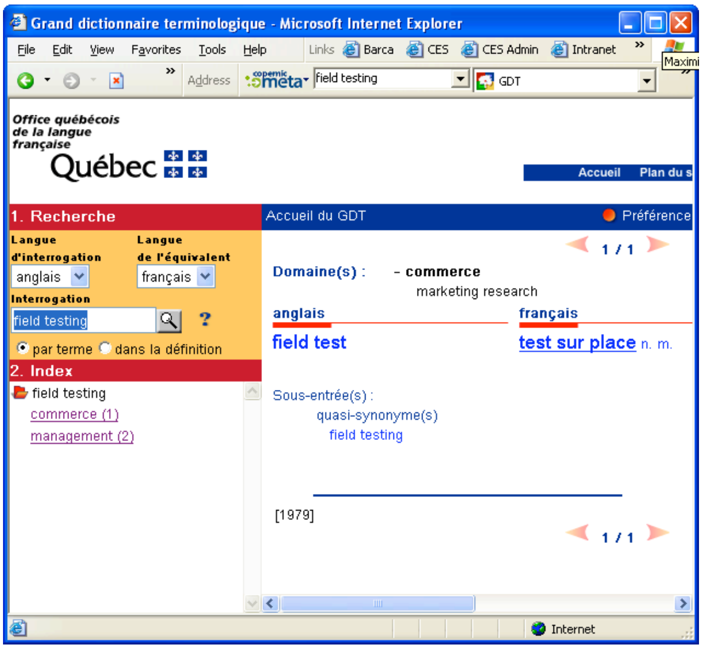
Experiment 1 & 2 Results

	Accuracy (%)	
baseline experiment	20.21	
hansard	26.11	
cisti individual	36.50	
cisti global	39.54	
cisti global + correction	41.26	†

† statistically significant at the 95% level

Experiment 3

- Terminological resource: the *Grand Dictionnaire Terminologique* (GDT)
- Try « exact match » translation
- Use the first entry



Grand dictionnaire terminologique - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links Barca CES CES Admin Intranet
Address
copernic meta
field testing
GDT
Maximi
Office québécois de la langue française
Québec
Accueil Plan du s
1. Recherche
Accueil du GDT
Préférence
2. Index
1 / 1
Internet

Langue d'interrogation: anglais

Langue de l'équivalent: français

Interrogation: field testing

● par terme ○ dans la définition

field testing
commerce (1)
management (2)

Domaine(s) : - commerce
marketing research

anglais	français
field test	test sur place n. m.

Sous-entrée(s) :
quasi-synonyme(s)
field testing

[1979]

Experiment 3

- **Important problem**: lot of author keyphrases are plural while the ressource contains only singular terms
- We handled limited inflectional morphology:
  - Detection of English plural (e.g.: word ending [^f]ves)
  - Singularization of English term (e.g.: ves → f)
  - Pluralization of French term (e.g: al → aux)

  - Simple heuristics for multi-word expressions
    ex.: pomme de terre → pomme**s** de terre
    pomme de terre frite → pomme**s** de terre frite**s**

Experiment 3

- How to choose a candidate in a terminological resource?
- Problem: different *domain,* different *translation*

  e.g: marché = market (finance)
  marché = market place (commerce)
  marché = contract (law)

- We introduce the Minimal Domain Set (MDS) algorithm

- **Idea**: find the minimal set of *coherent* domains covering all keyphrases.

Experiment 3

1. Calculate the frequency of each domain *F(D)*
2. Calculate the number of domain per keyphrase $|D_{Ki}|$
3. Sort Keyphrases in ascending order of $|D_{Ki}|$

   For each keyphrase:

4. **Likeliness**: From the list of domains D_{Ki}, build a reduced list containing only the domains with the highest frequency F(D)
5. **Coherence**: From this reduced list, select the domain which has the highest coherence with a member of MDS. Add this domain to MDS.

Experiment 3

bar (metallurgy, textile, law, automotive, …30 other!)

iron (metallurgy, shoe)

cement (shoe, textile)

rubber (shoe, rubber, leasure, graphic)

lace (clothing, shoe, brewing)

(Step 1) F(shoe) = 4
F(metallurgy) = 2
F(textile) = 2
F(*other*) = 1

(Step 2) $|D_{iron}| = 2$
$|D_{cement}| = 2$
$|D_{lace}| = 3$
$|D_{rubber}| = 4$
$|D_{bar}| = 34$

(Step 3) L := iron, cement, lace, rubber, bar.

Experiment 3

(first loop) **Iron (**metallurgy, shoe**)**

(Step 4) Domain likeliness: F(shoe) = 4

MDS = {shoe}

…

(last loop) **bar** (metallurgy, textile, law, automotive, …30 other!)

(Step 4) Domain likeliness: F(metallurgy) = F(textile) = 2

(Step 5) Domain coherence:
C(shoe, metallurgy) = 0.1, C(shoe, textile) = 0.3

MDS = {shoe, textile}

Experiment 3 Results

	Accuracy (%)	
GDT	35.66	†
GDT + morphology	38.30	
GDT + morph + MDS	39.22	‡

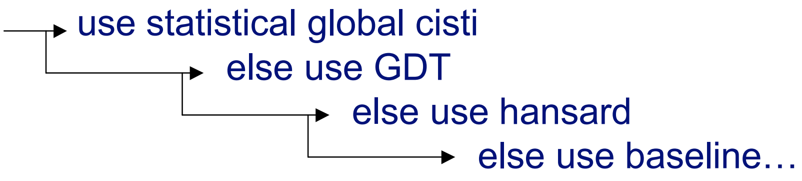
† statistically significant at the 99% level

‡ statistically significant at the 95% level

Experiment 4

Combining all previous modules:

- Cascading modules
  - For instance:



use statistical global cisti

else use GDT

else use hansard

else use baseline…

Experiment 4

- Combine modules (And optimize by genetic search)
- Use **all candidates** from phrase tables and GDT
- We used a weighting model involving six features. Features modify the original probabilities.
  - [1.071] candidate proposed by MT cisti model;
  - [0.227] candidate proposed by MT hansard model;
  - [0.477] candidate proposed by GDT;
  - [1.464] candidate proposed by more than one source;
  - [0.853] candidate using morphology;
  - [1.257] GDT candidate(s) in MDS;

Experiment 4 results

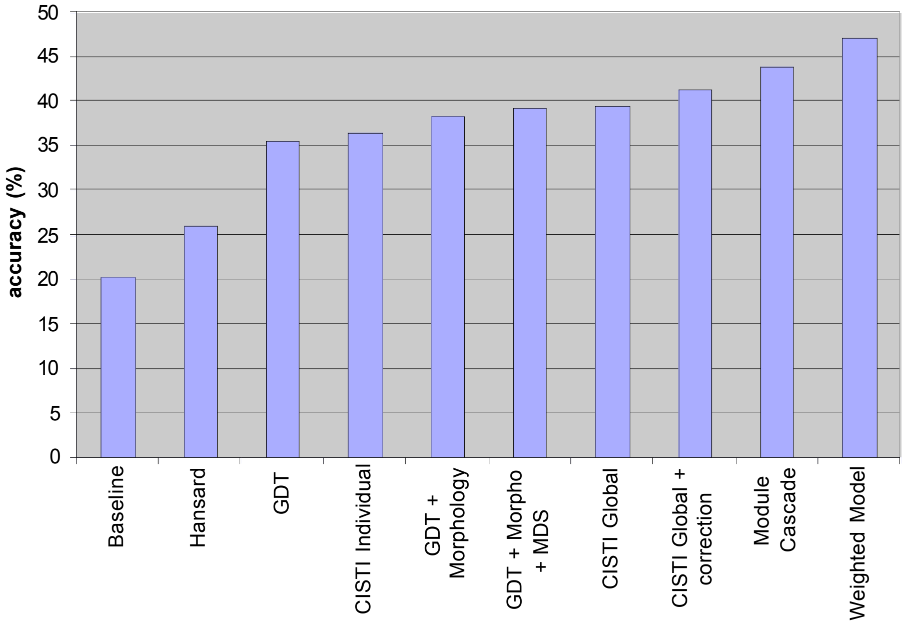
	Accuracy (%)
Module cascade	43.83
Best weighted model	47.16

Conclusion

- Overall best solution found by genetic search and 6 features
- Statistical MT and terminological resource are complementary
- We show how to use the keyphrase list to « disambiguate the sense » of a keyphrase

Conclusion

All experiments



Future Work

Two promising avenues:

- Include Termium, another Fr-En terminological ressource
- Follow (Jayaraman and Lavie, 2005) idea to combine module outputs and *create* candidate translations

NRC CNRC
Institute for
Information
Technology

Science
at work for
Canada

National Research
Council Canada

Conseil national
de recherches Canada

Canada