s

Improvements in Analogical Learning: Application to Translating multi-Terms of the Medical Domain

Philippe Langlais[1], François Yvon[2] and Pierre Zweigenbaum[2]

(1) DIRO
Univ. of Montreal
Montreal, Canada

(2) LIMSI-CNRS
Univ. Paris-Sud XI
Orsay, France

ITI-CNRC — Gatineau — June 12th, 2009

Motivations

- The blooming of new terms puzzles Machine Translation
  - current solution : **identifying** translations in parallel (Deléger et al., 2006) or comparable corpora (Morin et al., 2007 ; Chao & Zweigenbaum, 2002)
- Recent interest in Analogical Learning (AL)
  - as a fully-fledged translation engine (Lepage & Denoual, 2005)
  - as a device for translating unknown words (Langlais & Patry, 2007 ; Denoual, 2007)
  - as a mean to acquire morphological knowledge (Stroppa & Yvon, 2005 ; Hathout, 2006)
  - as a way of acquiring similarity of semantic relations between words (Turney & Littman, 2005, Turney, 2006)

Issues we wanted to address :

- Tackling practical issues in AL
- Comparing AL and SMT translations on the task of translating medical terms (small training set)

Overview

Motivations

Analogical Learning
- Formal Analogies
- Principle

Practical Issues
- Solver
- Search
- Over-generation

Experiments
- corpus
- metrics

Ongoing work

Motivations

Analogical Learning

- Formal Analogies
- Principle

Practical Issues

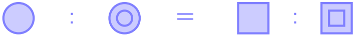
- Solver
- Search
- Over-generation

Experiments

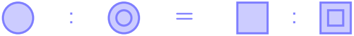
- corpus
- metrics

Ongoing work

Analogy

- Analogy : $[x : y = z : t] \equiv$ “x *is to* y *as* z *is to* t”
  - $[\mathit{mason} : \mathit{stone} = \mathit{carpenter} : \mathit{wood}]$ (Turney, 2006)
  -  (Lepage, 1998)

Analogy

- Analogy : $[x : y = z : t] \equiv$ "x *is to* y *as* z *is to* t"
  - [*mason* : *stone* = *carpenter* : *wood*] (Turney, 2006)
  -  (Lepage, 1998)

- Formal Analogy : analogy between forms
  - [*reader* : *unreadable* = *doer* : *undoable*] (Lepage, 1998)
  - [*keras* : *mengeraskan* = *kena* : *mengenakan*] (Lepage, 1998)

Some definitions of a formal analogy

- (Pirrelli & Yvon, 1999)

$$[x : y = z : t] \iff \text{or} \begin{cases} x = bc, y = bd, z = ac, t = ad \\ x = bc, y = ac, z = bd, t = ad \end{cases}$$

  - $[dream : dreamer = eat : eater]$
  - $[steal : ceal = stage : cage]$

- (Lepage, 1998)

$$[x : y = z : t] \Rightarrow \left[\begin{array}{lcl} \sigma(y,t) & = & -|x| + |y| + \sigma(x,z) \\ \sigma(z,t) & = & -|x| + |z| + \sigma(x,y) \\ \sigma(x,y,z,t) & = & -|x| + \sigma(x,y) + \sigma(x,z) \\ |t|_a & = & -|x|_a + |y|_a + |z|_a \; \forall a \end{array} \right.$$

  - $[believer : unbelievable = dreamer : undreamable]$

Definition (Stroppa & Yvon, 2005)

- **Def. :** $[x : y = z : t]$ **iff** we can find **factorizations** f_x, f_y, f_z and f_t such that, $\forall i \in [1, d]$:

$$(f_y^{(i)}, f_z^{(i)}) \in \left\{(f_x^{(i)}, f_t^{(i)}), (f_t^{(i)}, f_x^{(i)})\right\}$$

  - $f_x^{(i)}, f_y^{(i)}, f_z^{(i)}$ and $f_t^{(i)}$ are called the **factors**
  - the smallest d for which this holds is called the **degree**

- [*this guy drinks too much* : *this boat sinks* = *these guys drank too much* : *these boats sank*] because :

x	≡	this	guy	ϵ	dr	inks	too much
y	≡	this	boat	ϵ	s	inks	ϵ
z	≡	these	guy	s	dr	ank	too much
t	≡	these	boat	s	s	ank	ϵ

Definition (Stroppa & Yvon, 2005)

- **Def. :** $[x : y = z : t]$ **iff** we can find **factorizations** f_x, f_y, f_z and f_t such that, $\forall i \in [1, d]$:

$$(f_y^{(i)}, f_z^{(i)}) \in \left\{ (f_x^{(i)}, f_t^{(i)}), (f_t^{(i)}, f_x^{(i)}) \right\}$$

  - $f_x^{(i)}, f_y^{(i)}, f_z^{(i)}$ and $f_t^{(i)}$ are called the **factors**
  - the smallest d for which this holds is called the **degree**

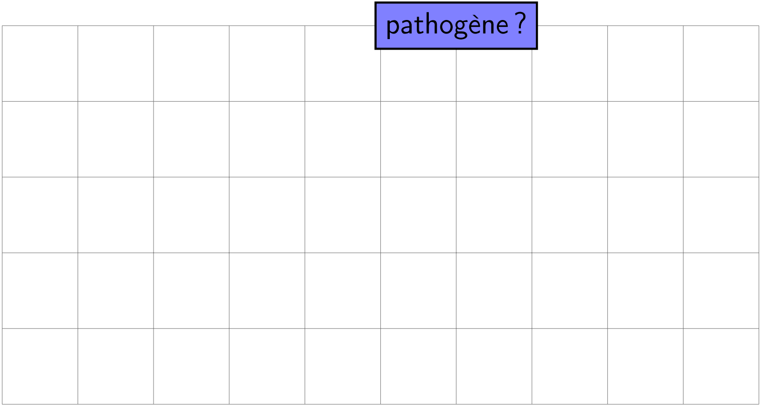
- [*this guy drinks too much* : *this boat sinks* = *these guys drank too much* : *these boats sank*] because :

x	$\equiv$	this	guy	ϵ	dr	inks	too much
y	$\equiv$	this	boat	ϵ	s	inks	ϵ
z	$\equiv$	these	guy	s	dr	ank	too much
t	$\equiv$	these	boat	s	s	ank	ϵ

  - the degree of this analogy is 6

Analogical Learning : Illustration

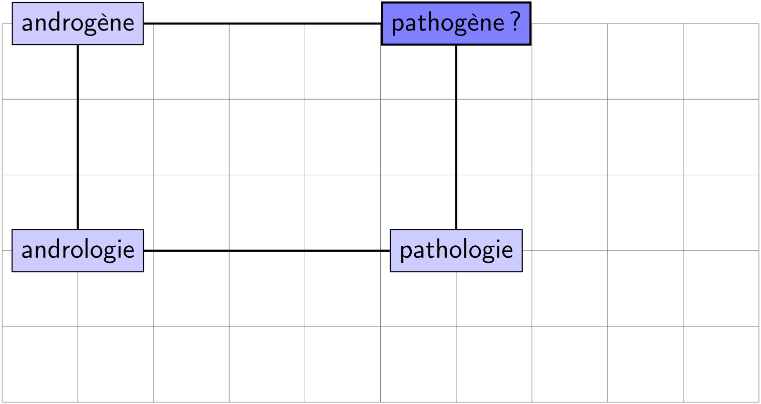
$$\mathcal{L} = \{\langle \textit{méthodologie}, \textit{methodology} \rangle, \langle \textit{angiolyse}, \textit{angiolysis} \rangle, \ldots\}$$



pathogène ?

Illustration (generator)

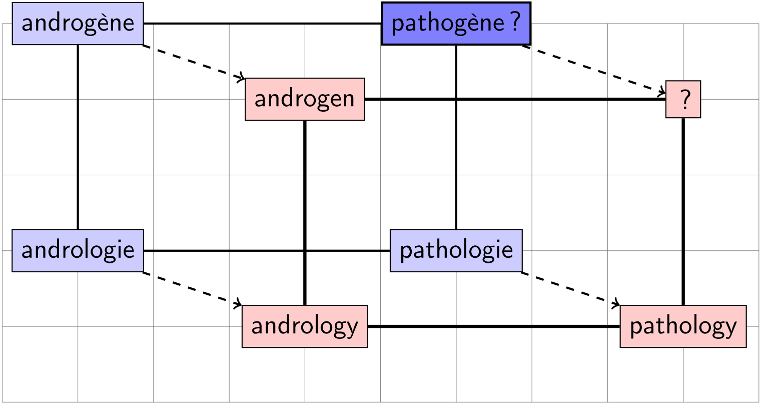
$\mathcal{L} = \{\langle \textit{méthodologie}, \textit{methodology}\rangle, \langle \textit{angiolyse}, \textit{angiolysis}\rangle, \ldots\}$



- Step 1 : find **source** analogies

Illustration (generator)

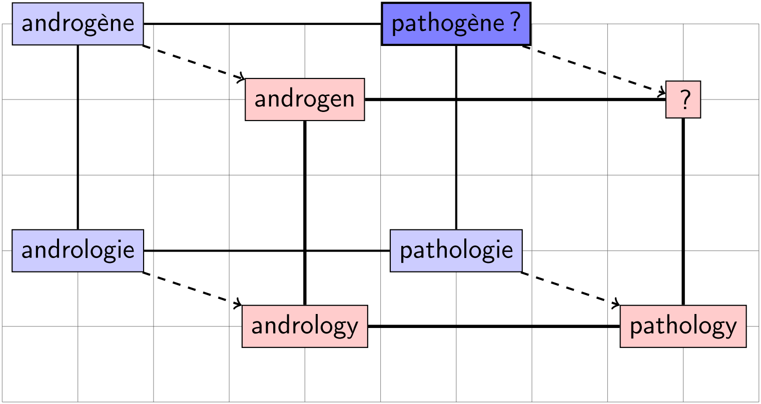
$\mathcal{L} = \{\langle \textit{méthodologie}, \textit{methodology} \rangle, \langle \textit{angiolyse}, \textit{angiolysis} \rangle, \ldots\}$



- Step 2 : Solve the **target** analogical equation

Illustration (generator)

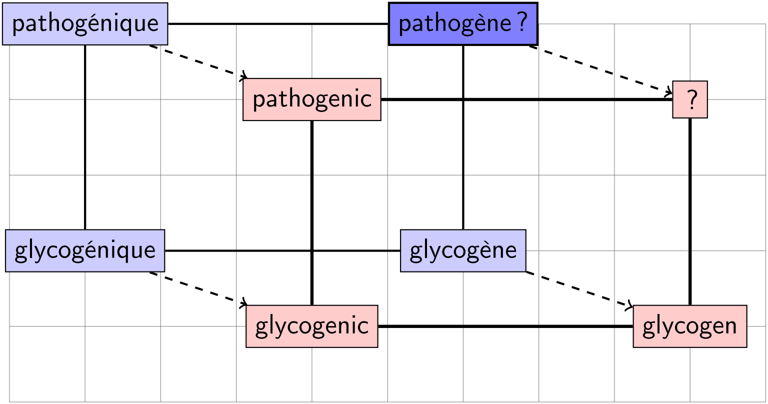
$\mathcal{L} = \{\langle \textit{méthodologie, methodology} \rangle, \langle \textit{angiolyse, angiolysis} \rangle, \ldots\}$



- *generated :* ϕ
- ? $\equiv$ pathogen, genpatho, ogpathen, pagthoen, paogthen, . . .

Illustration (generator)

$\mathcal{L} = \{\langle \textit{méthodologie}, \textit{methodology}\rangle, \langle \textit{angiolyse}, \textit{angiolysis}\rangle, \ldots\}$



- *generated : pathogen, genpatho, ogpathen, pagthoen, paogthen, . . .*
- ? ≡ pathogen, patoghen, opgathen, pathoegn, pathgeno, . . .

Illustration (generator)

- test term : pathogène
- 147 source analogies found in the training material
- 18 of the 3788 forms generated
  (a candidate translation can be generated by different analogies)

(pathogenic,43)	(pathogenous,34)	(ogenpathous,34)
(ogpathenous,33)	(genoupathos,33)	(genouspatho,33)
(ogenopathus,33)	(ogenoupaths,33)	(ogenouspath,33)
(ogenupathos,33)	(ogenuspatho,33)	(ogepathnous,33)
(genopathous,32)	(genpathoous,32)	(gepathonous,32)
(opathgenous,32)	(pathogen, 31)	(pathoogenus,31)
	. . .	

Illustration (selector)

(pathogenic,43)	(pathogenous,34)	~~(ogenpathous,34)~~
~~(ogpathenous,33)~~	~~(genoupathos,33)~~	~~(genouspatho,33)~~
~~(ogenopathu,33)~~	~~(ogenoupaths,33)~~	~~(ogenouspath,33)~~
~~(ogenupathos,33)~~	~~(ogenuspatho,33)~~	~~(ogepathnous,33)~~
~~(genopathous,32)~~	~~(genpathoous,32)~~	~~(gepathonous,32)~~
~~(opathgenous,32)~~	(pathogen,31)	~~(pathoogenus,31)~~
	...	

- Step 3 : Remove unlikely candidates

Motivations

Analogical Learning
- Formal Analogies
- Principle

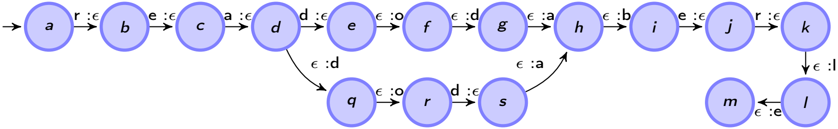
Practical Issues
- Solver
- Search
- Over-generation

Experiments
- corpus
- metrics

Ongoing work

Solvers

- We can built a finite-state **transducer** which produces the solutions to $[x : y = z : \ ?]$ while recognizing the form x (Yvon et al., 2003)



$$[reader : readable = doer : \ ?]$$

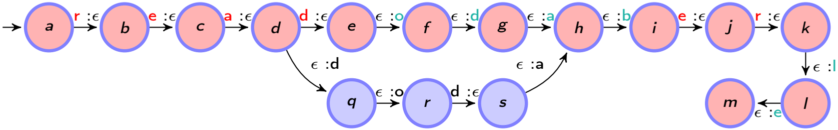
  - **problem** : building this automaton can face combinatorial problems

- We proposed a simple yet efficient way to sample this automaton

s	*nb*	$[reader : readable = doer : \ ?]$		
10	11	(doable,7)	(dabloe,3)	(adbloe,3)
10^2	22	(doable,28)	(dabloe,21)	(abldoe,21)
10^3	29	(doable,333)	(dabloe,196)	(abldoe,164)

Solvers

- We can built a finite-state **transducer** which produces the solutions to $[x : y = z : ?]$ while recognizing the form x (Yvon et al., 2003)



$[reader : readable = doer : ?] \Rightarrow$ odable

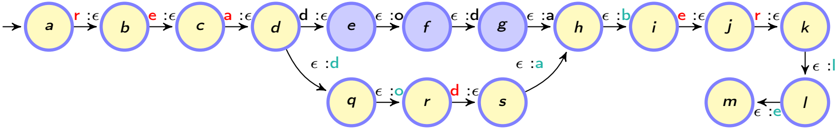
  - **problem** : building this automaton can face combinatorial problems

- We proposed a simple yet efficient way to sample this automaton

s	*nb*	$[reader : readable = doer : ?]$		
10	11	(doable,7)	(dabloe,3)	(adbloe,3)
10^2	22	(doable,28)	(dabloe,21)	(abldoe,21)
10^3	29	(doable,333)	(dabloe,196)	(abldoe,164)

Solvers

- We can built a finite-state **transducer** which produces the solutions to $[x : y = z : ?]$ while recognizing the form x (Yvon et al., 2003)



$[\textit{reader} : \textit{readable} = \textit{doer} : ?] \Rightarrow$ odable, **doable**

  - **problem** : building this automaton can face combinatorial problems

- We proposed a simple yet efficient way to sample this automaton

s	*nb*	$[\textit{reader} : \textit{readable} = \textit{doer} : ?]$		
10	11	(doable,7)	(dabloe,3)	(adbloe,3)
10^2	22	(doable,28)	(dabloe,21)	(abldoe,21)
10^3	29	(doable,333)	(dabloe,196)	(abldoe,164)

Search issues

We must find source triplets $(x, y, z) \in \mathcal{I}^3$ that define with t an analogy.

- brute-force : $o(|\mathcal{I}|^3)$ analogies to check

Search issues

We must find source triplets $(x, y, z) \in \mathcal{I}^3$ that define with t an analogy.

- brute-force : $o(|\mathcal{I}|^3)$ analogies to check
- turned into a quadratic number of equation solving (Lepage & Denoual, 2005) :

  1. consider $(x, y) \in \mathcal{I}^2$
  2. solve $[y : x = t :\ ?]$
  3. filter in the solutions z that belong to $\mathcal{I}$

  $\Rightarrow$ they define the triplets (x, y, z)

Follows from the property : $[x : y = z : t] \Leftrightarrow [y : x = t : z]$

Search issues

We must find source triplets $(x, y, z) \in \mathcal{I}^3$ that define with t an analogy.

- brute-force : $o(|\mathcal{I}|^3)$ analogies to check
- turned into a quadratic number of equation solving (Lepage & Denoual, 2005) :

  1. consider $(x, y) \in \mathcal{I}^2$
  2. solve $[y : x = t : ?]$
  3. filter in the solutions z that belong to $\mathcal{I}$

  $\Rightarrow$ they define the triplets (x, y, z)

  Follows from the property : $[x : y = z : t] \Leftrightarrow [y : x = t : z]$

- Still impractical for (not too) large input spaces
  $\Rightarrow$ sample pairs (x, y) (Langlais & Patry, 2007)

  1. define a neighborhood function $\mathcal{N}$ (thresholded edit-distance)
  2. sample x from $\mathcal{N}(t)$
  3. sample y from $\mathcal{N}(x)$

Search issues

- **Prop.** $[x : y = z : t] \Rightarrow |x|_c + |t|_c = |y|_c + |z|_c \ \forall c \in \mathcal{A}$ (Lepage,1998)
- We can find efficiently (x, y, z, t) such that $|x|_c + |t|_c = |y|_c + |z|_c \ \forall c \in \mathcal{A}$ (Langlais & Yvon, 2008)
- Our search strategy :
  1. consider all $x \in \mathcal{I}$
  2. search for all the pairs (y, z) satisfying the count property
  3. check for **true analogies**
     algorithm in $o(|x| \times |y| \times |z| \times |t|)$ *proposed by (Stroppa, 2005)*

Impact of the search-strategy

- **Task :** identifying in $\mathcal{I}$ the analogies of 1 000 word-forms

	a	%	(s)	a	%	(s)	a	%	(s)
our solution	34	83.1	0.2	261	94.1	0.5	746	96.4	1.2
Langlais & Patry	17	71.7	7.4	46	85.0	7.6	56	88.9	6.3
$\|\mathcal{I}\|$	20 000			50 000			84 076		

- a : average number of analogies per test form
- % : percentage of forms with at least one analogy found (coverage)
- (s) : average time (in seconds) spent per form

Impact of the search-strategy

- **Task :** identifying in $\mathcal{I}$ the analogies of 1 000 word-forms

	a	%	(s)	a	%	(s)	a	%	(s)
our solution	34	83.1	0.2	261	94.1	0.5	746	96.4	1.2
Langlais & Patry	17	71.7	7.4	46	85.0	7.6	56	88.9	6.3
$\lvert\mathcal{I}\rvert$	20 000			50 000			84 076		

- a : average number of analogies per test form
- % : percentage of forms with at least one analogy found (coverage)
- (s) : average time (in seconds) spent per form

Dealing with over-generation

- The generator produces many *(thousands)* target forms per source ones. . .
- Several solutions proposed :
  - filtering by frequency
    (Lepage & Denoual, 2005 ; Stroppa & Yvon, 2005 ; Denoual, 2007)
  - filtering forms unseen in a *(large)* set of *(target)* forms
    (Langlais & Patry, 2007)
  - filtering forms containing character-ngrams unseen in the training material
    (Lepage & Lardilleux, 2007)
  - learning to recognize meaningful **examples** from bad ones
    (our solution)

Supervised learning of good examples (on dev)

[*andrologie* : *pathologie* = *androgène* : *pathogène*]
[*andrology* : *pathology* = *androgen* : *paogthen*] ☹

⋮

[*otologiste* : *pathologiste* = *otogène* : *pathogène*]
[*otologist* : *pathologist* = *otogenic* : *pathogenic*] ☺

- 1000 terms of dev ⇒ ∼ 3 M. of examples ; ∼ 4 000 positive ones only
- Features used :
  - degree of the source and target analogies,
  - frequency of a candidate translation,
  - character-based ngram probabilities given to a candidate translation,
  - code-books of factors involved,
  - etc.

Selector

- voted-perceptron (Freund & Schapire, 1999)
  - 20 epochs
  - we removed examples which solution is frequent less than 3 times *(loss 3.4%)*
  - we trained many different feature representations
- task : identifying **positive** examples *(less than 1% of the examples)*
  - `s-best` : best voted-perceptron on dev
  - `argmax-f1` : pick to most frequent solution

(FI→EN)	p	r
`argmax-f1`	41.3	56.7
`s-best`	53.6	61.3

- systematic gains of the classifier in precision and recall over `argmax-f1`

Motivations

Analogical Learning
- Formal Analogies
- Principle

Practical Issues
- Solver
- Search
- Over-generation

Experiments
- corpus
- metrics

Ongoing work

Corpora

- Data extracted from the Medical Subject Headings (MeSH) thesaurus

	train			test	dev	test
f	nb	$u_f\%$	nb	$u_f\%$	$u_f\%$	oov%
FI	19 787	63.7	1 000	64.2	64.0	5.7
FR	17 230	29.8	1 000	30.8	28.3	36.3
RU	21 407	38.6	1 000	38.5	40.2	44.4
SP	19 021	31.1	1 000	31.7	33.3	36.6
SW	17 090	67.9	1 000	67.4	67.9	68.4

  - $u_f\%$ percentage of uni-terms in the *Foreign* part.
- Ex :
  - *speech articulation tests* ↔ *ääntämiskokeet* EN↔FI
  - *ovulation prediction* ↔ *ägglossningsförutsägelse* EN↔SW
  - *ischemic attack, transient* ↔ *accident ischémique transitoire* EN↔FR
  - *dentin-bonding agents* ↔ *agentes de recubrimiento dental adhesivo* EN↔SP
  - *ophtalmodynamometry* ↔ ОФТАЛЬМОДИНАМОМЕТРИЯ EN↔RU

Corpora

- Data extracted from the Medical Subject Headings (MeSH) thesaurus

	train			test	dev	test
f	*nb*	u_f%	*nb*	u_f%	u_f%	oov%
FI	19 787	63.7	1 000	64.2	64.0	5.7
FR	17 230	29.8	1 000	30.8	28.3	36.3
RU	21 407	38.6	1 000	38.5	40.2	44.4
SP	19 021	31.1	1 000	31.7	33.3	36.6
SW	17 090	67.9	1 000	67.4	67.9	68.4

  - u_f% percentage of uni-terms in the *Foreign* part.

- Ex :
  - *speech articulation tests* ↔ *ääntämiskokeet* EN↔FI
  - *ovulation prediction* ↔ *ägglossningsförutsägelse* EN↔SW
  - *ischemic attack, transient* ↔ *accident ischémique transitoire* EN↔FR
  - *dentin-bonding agents* ↔ *agentes de recubrimiento dental adhesivo* EN↔SP
  - *ophtalmodynamometry* ↔ ОФТАЛЬМОДИНАМОМЕТРИЯ EN↔RU

Corpora

- Data extracted from the Medical Subject Headings (MeSH) thesaurus

	train			test	dev	test
f	*nb*	u_f%	*nb*	u_f%	u_f%	oov%
FI	19 787	63.7	1 000	64.2	64.0	5.7
FR	17 230	29.8	1 000	30.8	28.3	36.3
RU	21 407	38.6	1 000	38.5	40.2	44.4
SP	19 021	31.1	1 000	31.7	33.3	36.6
SW	17 090	67.9	1 000	67.4	67.9	68.4

  - u_f% percentage of uni-terms in the *Foreign* part.

- Ex :
  - *speech articulation tests* ↔ *ääntämiskokeet* EN↔FI
  - *ovulation prediction* ↔ *ägglossningsförutsägelse* EN↔SW
  - *ischemic attack, transient* ↔ *accident ischémique transitoire* EN↔FR
  - *dentin-bonding agents* ↔ *agentes de recubrimiento dental adhesivo* EN↔SP
  - *ophtalmodynamometry* ↔ ОФТАЛЬМОДИНАМОМЕТРИЯ EN↔RU

Metrics

A.	*pleuropneumoni, smittsam*	1.	(pleuropneumonia, infectious,68)
		2.	(pleuropneumonia, contagious,28)
B.	*äggimplantation, försenad*	1.	(embryo implantation, delayed,22)
C.	*dragant*		

- **Coverage** the fraction of input forms for which the system can generate translations. If N_t words receive translations among N, then :

  $Cov = N_t/N$ $\quad N = 3,\ N_t = 2 \Rightarrow Cov = 2/3$

- **Recall at rank** k is the proportion of the N input forms for which a correct translation is output among the k first translations :

  $R_k = N_k/N$ $\quad R_1 = 1/3,\ R_2 = 2/3$

- **Precision at rank** k : proportion of forms for which a correct translation is output. Let N_k be the number of forms with the reference translation in the k first proposed. then :

  $P_k = N_k/N_t$ $\quad P_1 = 1/2,\ P_2 = 2/2 = 1$

Coverage

	FI	FR	RU	SP	SW
EN →	47.1	41.2	46.2	47.0	42.8
EN ←	44.8	38.5	42.1	42.6	44.6

- Less than half of the test terms received a translation by the analogical device . . .
- With a training material 3 times larger, we measured a huge increase in coverage : 73.4% (sp2en) 79.7% (en2sp)

Precision & Recall

		FI→EN		FR→EN		RU→EN		SP→EN		SW→EN	
	k	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*
argmax-f	1	41.3	17.3	46.7	16.8	47.8	18.6	48.7	19.2	43.4	18.1
s-best	1	53.5	20.8	56.9	19.3	58.5	20.3	63.2	22.5	50.4	21
oracle	1	100	30.5	100	26.3	100	28.5	100	30.6	100	29.5
argmax-f	10	61.6	25.8	62.8	22.6	61.7	24.0	69.3	27.3	62.1	25.9
s-best	10	69.4	27.0	69.0	23.4	71.8	24.9	78.4	27.9	65.7	27.4

- between 19.3% and 22.5% of the test terms translated with a precision ranging from 50.4% to 63.2%
- oracle : a perfect selector

Precision & Recall

	k	FI→EN		FR→EN		RU→EN		SP→EN		SW→EN	
		P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*
argmax-f	1	41.3	17.3	46.7	16.8	47.8	18.6	48.7	19.2	43.4	18.1
s-best	1	53.5	20.8	56.9	19.3	58.5	20.3	63.2	22.5	50.4	21
oracle	1	100	30.5	100	26.3	100	28.5	100	30.6	100	29.5
argmax-f	10	61.6	25.8	62.8	22.6	61.7	24.0	69.3	27.3	62.1	25.9
s-best	10	69.4	27.0	69.0	23.4	71.8	24.9	78.4	27.9	65.7	27.4

- between 19.3% and 22.5% of the test terms translated with a precision ranging from 50.4% to 63.2%
- oracle : a perfect selector

Precision & Recall

		FI→EN		FR→EN		RU→EN		SP→EN		SW→EN	
	k	Pk	Rk	Pk	Rk	Pk	Rk	Pk	Rk	Pk	Rk
`argmax-f`	1	41.3	17.3	46.7	16.8	47.8	18.6	48.7	19.2	43.4	18.1
`s-best`	1	53.5	20.8	56.9	19.3	58.5	20.3	63.2	22.5	50.4	21
oracle	1	100	30.5	100	26.3	100	28.5	100	30.6	100	29.5
`argmax-f`	10	61.6	25.8	62.8	22.6	61.7	24.0	69.3	27.3	62.1	25.9
`s-best`	10	69.4	27.0	69.0	23.4	71.8	24.9	78.4	27.9	65.7	27.4

- between 19.3% and 22.5% of the test terms translated with a precision ranging from 50.4% to 63.2%
- `oracle` : a perfect selector

Precision & Recall

		FI→EN		FR→EN		RU→EN		SP→EN		SW→EN	
	k	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*	P*k*	R*k*
`argmax-f`	1	41.3	17.3	46.7	16.8	47.8	18.6	48.7	19.2	43.4	18.1
`s-best`	1	53.5	20.8	56.9	19.3	58.5	20.3	63.2	22.5	50.4	21
oracle	1	100	30.5	100	26.3	100	28.5	100	30.6	100	29.5
`argmax-f`	10	61.6	25.8	62.8	22.6	61.7	24.0	69.3	27.3	62.1	25.9
`s-best`	10	69.4	27.0	69.0	23.4	71.8	24.9	78.4	27.9	65.7	27.4

- between 19.3% and 22.5% of the test terms translated with a precision ranging from 50.4% to 63.2%
- `oracle` : a perfect selector

Precision & Recall

		FI→EN		FR→EN		RU→EN		SP→EN		SW→EN	
	k	Pk	Rk	Pk	Rk	Pk	Rk	Pk	Rk	Pk	Rk
`argmax-f`	1	41.3	17.3	46.7	16.8	47.8	18.6	48.7	19.2	43.4	18.1
`s-best`	1	53.5	20.8	56.9	19.3	58.5	20.3	63.2	22.5	50.4	21
oracle	1	100	30.5	100	26.3	100	28.5	100	30.6	100	29.5
`argmax-f`	10	61.6	25.8	62.8	22.6	61.7	24.0	69.3	27.3	62.1	25.9
`s-best`	10	69.4	27.0	69.0	23.4	71.8	24.9	78.4	27.9	65.7	27.4

- between 19.3% and 22.5% of the test terms translated with a precision ranging from 50.4% to 63.2%
- `oracle` : a perfect selector

Combining Analogical & SMT devices

- phrase-based SMT engine :
  - Pharaoh (Koehn, 2004), phrase-table and language model trained on train
  - 8 coefficients tuned on dev
  - **basic unit** : character (Vilar et al., 2007 ; Paul et al., 2009 ; Deselaers et al., 2009)
    - too many oov words ; small training corpus *(word-based SMT does not work)*
    - direct comparison of SMT and AL translation devices
  - on dev, bleu scores range from 67.2 (en2fi) to 77.0 (ru2en)

	→ EN		← EN	
	P_{smt}	ΔB	P_{smt}	ΔB
FI	20.2	+7.4	21.6	+6.4
FR	19.9	+5.3	17.0	+6.0
RU	24.1	+3.1	28.0	+6.4
ES	22.1	+4.9	26.4	+5.5
SW	25.9	+4.2	31.6	+3.2

- SMT : lower precision, but higher recall

Examples

sw	*aikakauslehdet aiheena*
ref	*periodicals as topic*
ana	*periodicals as topic*
smt	*timenancylages, topic*

sw	*alfasalpaajat*
ref	*adrenergic alpha-antagonists*
ana	*adrenergic alpha-antagonists*
smt	*alphablockers*

sp	*instituciones de atención ambulatoria*
ref	*ambulatory care facilities*
ana	*ambulatory care facilities*
smt	*institutions, atention ambulatory*

fi	*märkivä kilpirauhastulehdus*
ref	*thyroiditis, suppurative*
ana	*thyroiditis suppurativa*
smt	*rativa thyroid glandorum*

fr	*malformations de la machoîre*
ref	*jaw abnormalities*
ana	*jaw congenital abnormalities*
smt	*malformations jawory*

fi	*rasva-alkoholit*
ref	*fatty alcohols*
ana	*lipid alcohols*
smt	*fatty-alcohols*

Examples

sw	*aikakauslehdet aiheena*

ref	*periodicals as topic*
ana	*periodicals as topic*
smt	*timenancylages, topic*

sp	*instituciones de atención ambulatoria*

ref	*ambulatory care facilities*
ana	*ambulatory care facilities*
smt	*institutions, atention ambulatory*

fr	*malformations de la machoîre*

ref	*jaw abnormalities*
ana	*jaw congenital abnormalities*
smt	*malformations jawory*

sw	*alfasalpaajat*

ref	*adrenergic alpha-antagonists*
ana	*adrenergic alpha-antagonists*
smt	*alphablockers*

fi	*märkivä kilpirauhastulehdus*

ref	*thyroiditis, suppurative*
ana	*thyroiditis suppurativa*
smt	*rativa thyroid glandorum*

fi	*rasva-alkoholit*

ref	*fatty alcohols*
ana	*lipid alcohols*
smt	*fatty-alcohols*

Examples

sw	*aikakauslehdet aiheena*
ref	*periodicals as topic*
ana	*periodicals as topic*
smt	*timenancylages, topic*

sw	*alfasalpaajat*
ref	*adrenergic alpha-antagonists*
ana	*adrenergic alpha-antagonists*
smt	*alphablockers*

sp	*instituciones de atención ambulatoria*
ref	*ambulatory care facilities*
ana	*ambulatory care facilities*
smt	*institutions, atention ambulatory*

fi	*märkivä kilpirauhastulehdus*
ref	*thyroiditis, suppurative*
ana	*thyroiditis suppurativa*
smt	*rativa thyroid glandorum*

fr	*malformations de la machoîre*
ref	*jaw abnormalities*
ana	*jaw congenital abnormalities*
smt	*malformations jawory*

fi	*rasva-alkoholit*
ref	*fatty alcohols*
ana	*lipid alcohols*
smt	*fatty-alcohols*

Examples

sw	*aikakauslehdet aiheena*
ref	*periodicals as topic*
ana	*periodicals as topic*
smt	*timenancylages, topic*

sw	*alfasalpaajat*
ref	*adrenergic alpha-antagonists*
ana	*adrenergic alpha-antagonists*
smt	*alphablockers*

sp	*instituciones de atención ambulatoria*
ref	*ambulatory care facilities*
ana	*ambulatory care facilities*
smt	*institutions, atention ambulatory*

fi	*märkivä kilpirauhastulehdus*
ref	*thyroiditis, suppurative*
ana	*thyroiditis suppurativa*
smt	*rativa thyroid glandorum*

fr	*malformations de la machoîre*
ref	*jaw abnormalities*
ana	*jaw congenital abnormalities*
smt	*malformations jawory*

fi	*rasva-alkoholit*
ref	*fatty alcohols*
ana	*lipid alcohols*
smt	*fatty-alcohols*

Recap

- We proposed practical solutions to analogical learning :
  - a solver which finds more solutions than the one of (Lepage, 1998)
  - a fast and efficient search-procedure for identifying (source) analogies
  - a better way to identify spurious solutions than by frequency alone
- We applied these enhancements to translating multi-terms of the medical domain :
  - comparable performance over 10 translation directions
  - at best, we could translate 30% of the terms with a perfect precision
  - higher precision than a character-based SMT engine, but lower recall
  - a straightforward combination of AL + SMT leads to an absolute improvement of 5.3 Bleu points over the SMT alone.

Motivations

Analogical Learning
- Formal Analogies
- Principle

Practical Issues
- Solver
- Search
- Over-generation

Experiments
- corpus
- metrics

Ongoing work

Analogy & Morphologie

Lexique

- wijsneuzig
- eenponder
- bedrijfspsychologie
- breedtecirkel
- rudolf
- terrasland
- conventualis
- luchtbad
- sliding
- bajonetaanval
- operatieveld
- hamerspie

Mot : prozabewerking

Segmentation ?

Analogy & Morphologie

Lexique

wijsneuzig
eenponder
bedrijfspsychologie
breedtecirkel
rudolf
terrasland
conventualis
luchtbad
sliding
bajonetaanval
operatieveld
hamerspie

Mot : prozabewerking

Segmentation ?

$$[\textit{prozabewerking} : \textit{prozawerk} = \textit{betekening} : \textit{teken}]$$

$$\begin{array}{lccccc}
f_{prozabewerking} & \equiv & \textit{proza} & \textit{be} & \textit{werk} & \textit{ing} \\
f_{prozawerk} & \equiv & \textit{proza} & \epsilon & \textit{werk} & \epsilon \\
f_{betekening} & \equiv & \epsilon & \textit{be} & \textit{teken} & \textit{ing} \\
f_{teken} & \equiv & \epsilon & \epsilon & \textit{teken} & \epsilon
\end{array}$$

Analogy & Morphologie

Lexique

wijsneuzig
eenponder
bedrijfspsychologie
breedtecirkel
rudolf
terrasland
conventualis
luchtbad
sliding
bajonetaanval
operatieveld
hamerspie

Mot : prozabewerking

Segmentation ?

$[\textit{prozawerk} : \textit{invloed} = \textit{prozabewerking} : \textit{beinvloedin}$

$$\begin{array}{llllll}
f_{prozawerk} & \equiv & \textit{proza} & \textit{be} & \textit{werk} & \textit{ing} \\
f_{invloed} & \equiv & \textit{proza} & \epsilon & \textit{werk} & \epsilon \\
f_{prozabewerking} & \equiv & \epsilon & \textit{be} & \textit{teken} & \textit{ing} \\
f_{beinvloeding} & \equiv & \epsilon & \epsilon & \textit{teken} & \epsilon
\end{array}$$

Analogy & Morphology

EN (16) *f* *factorisation*	DE (26) *f* *factorisation*	NL (26) *f* *factorisation*
18 **in+dent+ation**	92 unerbittlich+keit	18 p+r+ozabewerking
11 indent+ation	26 une+r+bittlichkeit	16 **proza+be+werk+ing**
7 ind+entation	14 **un+er+bitt+lich+keit**	14 prozab+e+werking
7 inden+tation	12 un+e+rbittlichkeit	12 pr+o+zabewerking
4 in+den+tation	12 unerbitt+lichkeit	10 proz+a+bewerking

nbf rang	nbf rang	nbf rang
9.3 2.2	22.7 2.3	29.9 4.9

- liens entre analogie formelle et morphologie (Langlais, 2009)
- participation à MorphoChalenge 2009 (Jean-François Lavallé, MSc)

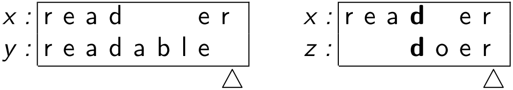
Thank you for your attention

Questions ?

Solvers

$$[\textit{reader} : \textit{readable} = \textit{doer} : \ ?]$$

- The solver of (Lepage,1998) :
  1. edit-distance computation (between x and y ; and between x and z)
  2. deterministic automaton
     - state : edit-operations at both cursors
     - action : copy one symbol from y or z into the solution ; move one or both cursors



x : r e a d e r
y : r e a d a b l e

x : r e a **d** e r
z : **d** o e r

  - **problem** : fortuitous alignments of symbols ⇒ dabloe

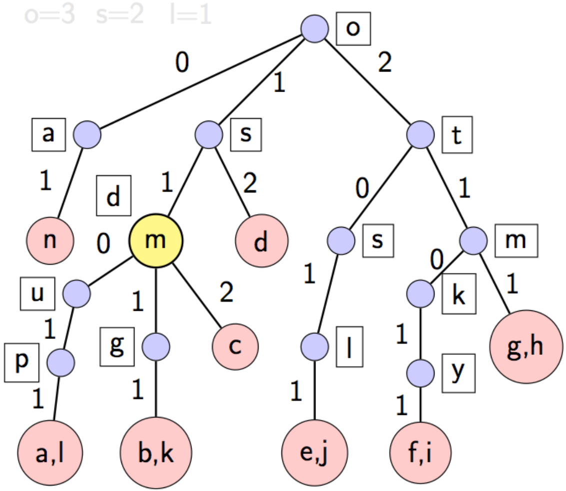
- We adapted the solver of (Stroppa & Yvon, 2005)

Generator

	Cov	P_1	R_1	P_{100}	R_{100}	R_∞
→ FI	**47.1**	*31.6*	14.9	*57.7*	27.2	31.9
FR	41.2	35.4	*14.6*	60.4	24.9	*26.5*
RU	46.2	40.5	18.7	69.9	32.3	34.8
ES	47.0	41.5	19.5	69.1	**32.5**	**35.9**
SW	42.8	36.0	15.4	66.8	28.6	31.9
← FI	44.8	36.6	16.4	66.7	29.9	33.2
FR	*38.5*	47.0	18.1	69.9	*26.9*	29.4
RU	42.1	**49.4**	**20.8**	70.3	29.6	32.3
ES	42.6	47.7	20.3	**75.1**	32.0	33.7
SW	44.6	40.8	18.2	69.5	31.0	32.9

- Coverage varies from 38.5% (fr2en) to 47.1% (en2fi)
- Recall (R_∞) is rather low : 26.5% (en2fr) to 39.5% (en2sp)
- On a much larger task (3 times more terms in the training material), we measured a huge increase in coverage : 73.4% (en2sp) 70.7% (en2sp)

A tree-count



a	**soup**
b	**gods**
c	**odds**
d	sos
e	solo
f	tokyo
g	moot
h	moto
i	kyoto
j	oslo
k	**dogs**
l	**opus**
m	**os**
n	a

A tree-count

- Input space : 11 317 717 forms
- Values averaged over 1 000 retrievals :

ratio	time (ms)	|*frontier*|	|nodes|
1/1000	5.5e-05	38	6.8
1/100	0.0003	150	6.3
1/10	0.003	1082	6.6
1/5	0.0055	1655	6.5
1/1	0.02	3921	5.8

- Memory and computation requirements roughly linear with the input space

Generalizing a phrase-table

$\lvert\mathcal{L}\rvert$	n		input s	input $\%s$	input (s)	output t	output $\%t$	output (s)
300t	10^3	rand	21	42.1	2	226	31.4	4
		ed	22	38.0	2	260	29.9	8
		ev	47	74.3	1	707	58.8	17
	∞		1046	77.2	206	10413	61.9	101
500t	10^3	rand	9	37.1	9	92	27.3	1
		ed	17	37.9	9	209	28.8	7
		ev	46	75.2	3	682	59.6	16
	∞		1155	81.5	3062	10856	65.1	108
11M	10^3	ev	48	76.4	11	743	76.0	19

Classifier versus most-frequent

	FI→EN		FR→EN		RU→EN		ES→EN		SW→EN	
	p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
`argmax-f1`	41.3	56.7	46.7	63.9	48.1	65.6	49.2	63.4	43.2	61.0
`s-best`	53.6	61.3	57.5	68.4	61.9	66.7	64.3	70.0	53.1	64.4

Analogie et Traduction statistique

Table de Segments

" asked the ||| demande le
" asked ||| " lui demanda
" asking the commission ||| " demande à la commission
" aspirin " , a ||| palliatif , elle constitue un
" aspirin " , ||| palliatif ,

Segment : a été discutée et

- espace d'entrée de plusieurs millions de formes . . .

Analogie et Traduction statistique

Table de Segments

```
" asked the ||| demande le
" asked ||| " lui demanda
" asking the commission ||| " demande à la commission
" aspirin " , a ||| palliatif , elle constitue un
" aspirin " , ||| palliatif ,
```

Segment : a été discutée et

Traduction ?

- espace d'entrée de plusieurs millions de formes . . .

Analogie et Traduction statistique

- âgées à leur sort. [1079]
  (old to die . ,57) (old on their own .,56) (old in the lurch .,53) (old to their fate .,41) (very old to die .,35) . . .
- ' acquis soient transposées [3610]
  (acquis are transposed,47) (acquis be transposed with,38) (acquis will be transposed,37) . . .
- a caractérisé la réunification allemande [3655]
  (has characterised of german reunification, ,24) (has characterised german reunification,20) . . .
- acceptables , sans mettre en [9985]
  (acceptable without calling into,23) (acceptable, without calling into,21) . . .
- a été discutée et [406223]
  (were debated and,151) (was discussed this and,133) (was discusseds thi and,123) (has been discussed and has,119). . .