Introduction à la traduction automatique (statistique)

felipe@iro.umontreal.ca

RALI
Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**

février 2015

Plan

rali

Crédits

- Chapitre 13 dans *[Manning and Schütze, 1999]*
- Chapitre 21 dans *[Jurafsky and Martin, 2000]*
- Tutoriel donné par Luong, Cho et Manning à ACL 2016
  https://sites.google.com/site/acl16nmt/

- certaines acétates prises à :
  - F. Yvon,
  - D. Déchelotte,
  - K. Knight & P. Koehn (tutoriel SMT),
  - H. Schwenk

rali

(Hutchins, 2005)

1949 Warren Weaver's (Rockfeller Foundation), théorie de l'information

(Hutchins, 2005)

1949 Warren Weaver's (Rockfeller Foundation), théorie de l'information

1960 Russe/Anglais, textes scientfiques et techniques
Initialement plutôt approches "classiques", IA + TAL : utilisation de Parseurs, de règles développées par des humains, . . .

(Hutchins, 2005)

- 1949 Warren Weaver's (Rockfeller Foundation), théorie de l'information
- 1960 Russe/Anglais, textes scientfiques et techniques
  Initialement plutôt approches "classiques", IA + TAL : utilisation de Parseurs, de règles développées par des humains, . . .
- 1966 Rapport ALPAC (Automatic Language Processing Advisory Committee)

(Hutchins, 2005)

1949 Warren Weaver's (Rockfeller Foundation), théorie de l'information

1960 Russe/Anglais, textes scientfiques et techniques
Initialement plutôt approches "classiques", IA + TAL : utilisation de Parseurs, de règles développées par des humains, . . .

1966 Rapport ALPAC (Automatic Language Processing Advisory Committee)

1970s Systran, système Meteo

(Hutchins, 2005)

- 1949 Warren Weaver's (Rockfeller Foundation), théorie de l'information
- 1960 Russe/Anglais, textes scientfiques et techniques
  Initialement plutôt approches "classiques", IA + TAL : utilisation de Parseurs, de règles développées par des humains, . . .
- 1966 Rapport ALPAC (Automatic Language Processing Advisory Committee)
- 1970s Systran, système Meteo
- 1990s Traduction statistique (IBM)

(Hutchins, 2005)

1949	Warren Weaver's (Rockfeller Foundation), théorie de l'information
1960	Russe/Anglais, textes scientfiques et techniques Initialement plutôt approches "classiques", IA + TAL : utilisation de Parseurs, de règles développées par des humains, . . .
1966	Rapport ALPAC (Automatic Language Processing Advisory Committee)
1970s	Systran, système Meteo
1990s	Traduction statistique (IBM)
2010s	Traduction neuronale (deep)

rali

Quelques faits

En vrac :

- fin 80 : traduction sur ordinateur personnel

Quelques faits

En vrac :

- fin 80 : traduction sur ordinateur personnel
- fin 90 : traduction sur la toile :
  - Alta Vista – Babel Fish – (Systran)
  - Google (initialement avec Systran)
  - Microsoft (Systran + TS à l'interne)
  - Language Weaver

Quelques faits

En vrac :

- fin 80 : traduction sur ordinateur personnel
- fin 90 : traduction sur la toile :
  - Alta Vista – Babel Fish – (Systran)
  - Google (initialement avec Systran)
  - Microsoft (Systran + TS à l'interne)
  - Language Weaver
- $$
  - 30% du budget du parlement européen
  - En 2004 : 1650 traducteurs professionnels employés à la Commission Européenne
  - 75% des pages internet sont monolingues (à voir . . .)
  - $\sim$ 5k langues, 68 langues parlées par + de 10M de personnes
    https://www.linguisticsociety.org/content/how-many-languages-are-there-world

rali

A recent UNESCO report on multilingualism states that languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society. From the very beginning, Europe had decided to keep its cultural and linguistic richness and diversity alive during the process of becoming an economic and political union. For maintaining the policy of multilingualism, the EU's institutions spend about one billion Euros a year on translating texts and interpreting spoken communication. For all European economies the translation costs for compliance with the laws and regulations are much higher.

Strategic Research Agenda for Multilingual Europe 2020, Springer 2013

*English, Mandarin, Punjabi, Vietnamese, French : These are, in descending order, the five most common languages spoken in **Vancouver**. The West Coast metropolis ranks as one of the most multilingual cities in the country. **Multilingualism is increasingly widespread in major cities—especially in Europe**, where high population density is conducive to interaction among cultural communities.*

The challenges of multilingualism in cities : An international research project, Gazette 5 oct. 2011

Many less resourced languages (LRL) that are thriving to get a place in the digital space and that could profit of the new opportunities offered by the Internet and digital devices will seriously face digital extinction if they are not supported by Language Technologies.

CFP LTC Workshop on "Less Resourced Languages, new technologies, new challenges and opportunities", 2013

Multilinguisme

- enjeux sociétaires :
  - *e-governance*
  - éducation
  - égalité sociale
  - etc.
- enjeu économique
- traduction automatique = technologies émergentes
  - domaine scientifique complexe faisant intervenir tous les aspects du traitement du langage naturel
  - pas de pratique encore cimentée : We need you !

rali

L'ordre des mots varie entre les langues
Belle marquise...

Exemples :

- Anglais :
  - IBM bought Lotus
  - Reporters said IBM bought Lotus
- Japonais :
  - IBM Lotus bought
  - Reporters IBM Lotus bought said
- Français :
  - une nouvelle voiture
  - une voiture nouvelle

Résolution des références

Il l'aime

systran : it likes it
google : he likes

Julie demande à Paul de ne plus la regarder

systran : Julie asks Paul more to look at it
google : Julie asks Paul no longer look

Julie demande à Paul de lui raconter une blague

systran : Julie asks Paul to tell him a joke
google : Julie asks Paul to tell a joke

Quelques problèmes de sémantique

Ambiguïté sémantique : multiplicité des sens d'un mot

- Anglais : plant (arbre ou entreprise) ; bank (banque ou bord d'une rivière)
- Français : allumer (une cigarette ou le moteur)
  couper (les cheveux (en 4) ou le moteur)

$\Rightarrow$ Souvent les sens différents correspondent à des traductions différentes

Idiomes

- Expressions poly-lexématiques qu'on ne peut traduire mot par mot (= non-compositionnelles)
- être au pied du mur $\rightarrow$ To be at the foot of the wall ?
- tenir sa langue $\rightarrow$ keep ones tongue ?
- ne pas mâcher ses mots $\rightarrow$ not to chew ones words ?

Problèmes de morpho-syntaxe

Utilisation des pronoms

- Certaine langues autorisent l'omission des pronoms (eg. espagnol, italien)
- Souvent la forme verbale détermine le bon pronom
- he, she ou it ?

  Atraversó$_v$ el río$_n$ flotando$_{pp}$ $\leftrightarrow$ it$_{pr}$ floated$_v$ across$_p$ the river

Problèmes de morpho-syntaxe

Utilisation des pronoms

- Certaine langues autorisent l'omission des pronoms (eg. espagnol, italien)
- Souvent la forme verbale détermine le bon pronom
- he, she ou it ?

  Atravesó$_v$ el río$_n$ flotando$_{pp}$ $\leftrightarrow$ it$_{pr}$ floated$_v$ across$_p$ the river

Problèmes de morpho-syntaxe

Utilisation des pronoms

- Certaine langues autorisent l'omission des pronoms (eg. espagnol, italien)
- Souvent la forme verbale détermine le bon pronom
- he, she ou it ?

  Atravers$ó_v$ el río$_n$ flotando$_{pp}$ $\leftrightarrow$ it$_{pr}$ floated_v across$_p$ the river

Marques flexionnelles

- He is nice $\rightarrow$ Il est beau vs She is nice $\rightarrow$ Elle est belle : accord d'un côté mais pas de l'autre

$\Rightarrow$ En général, la traduction est plus difficile quand la cible est morphologiquement plus riche que la source

Approches à la traduction vues dans ces acétates

Approches :

- Traduction par règles
  - Systran en a longtemps été l'archétype
- Approches utilisant des textes déjà traduits (traduction par l'exemple (EBMT))
  - "Expertise" contenue dans des traductions humaines
  - Minimise le problème d'acquisition de connaissances
    - approche statistique (par mots, par segments, autres)
    - approche neuronale

rali

Traduction Statistique

- la TS a fait une percée incroyable
- il est dit *ad libitum* que la TS s'améliore

  However, the overall user experience of automated translation services tends to be very poor, often including apparently nonsensical and obvious errors. Swapping common terms for non- equivalent terms in other languages as well as inverting sentence meaning are also common. Piron estimates that, at most, such machine translation will cater for about 25% of a translator's needs with the remaining 75% requiring human intervention to produce publishable quality translation.

  Lost in Translation – the challenges of multilingualism, Martin Flynn, 2011
- de très bons outils (Moses, Joshua, etc.)

rali

Traduction Neuronale

- percée spectaculaire
- https://www.theverge.com/2016/9/27/13078138/google-translate-ai-machine-learning-gnmt

> **Google's AI translation system is approaching human-level accuracy**
>
> *But there's still significant work to be done*
>
> By Nick Statt | @nickstatt | Sep 27, 2016, 2:07pm EDT
>
> SHARE

- de très bons outils (Open NMT, Nematus, etc.)

Le modèle du canal bruité *[Brown et al., 1993]*

Canal bruité

- f une phrase du langage source (french),
- e une phrase du langage cible (english),
- traduire $\Leftrightarrow$ résoudre :

$$\underset{e}{\operatorname{argmax}}\, P(e|f) = \underset{e}{\operatorname{argmax}}\, P(f|e)P(e)$$

Deux modèles

- $p(f|e)$ définit le *modèle de transfert*
- $p(e)$ définit le *modèle de langue*

Un décodeur

- problème NP-complet (Knight, 2001)

rali

Point de départ : un corpus parallèle

- Un corpus parallèle

• The Legislative Assembly convened at 3.30 pm. • Mr. Quirke (Clerk-Designate) : • THURSDAY, APRIL 1, 1999	• sitamiq, ipuru 1, 1999 • maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut • mista kuak (titiraqti - tikkuaqtausimajuq) :

Point de départ : un corpus parallèle

- Un corpus parallèle + *Aligneur* = **bitexte**

• The Legislative Assembly convened at 3.30 pm. • Mr. Quirke (Clerk-Designate) : • THURSDAY, APRIL 1, 1999	• sitamiq, ipuru 1, 1999 • maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut • mista kuak (titiraqti - tikkuaqtausimajuq) :

Point de départ : un corpus parallèle

- Un corpus parallèle + *Aligneur* = **bitexte**

• The Legislative Assembly convened at 3.30 pm. • Mr. Quirke (Clerk-Designate) : • THURSDAY, APRIL 1, 1999	• sitamiq, ipuru 1, 1999 • maligaliurvik matuiqtaulauqtuq 3 :30mi unnusakkut • mista kuak (titiraqti - tikkuaqtausimajuq) :

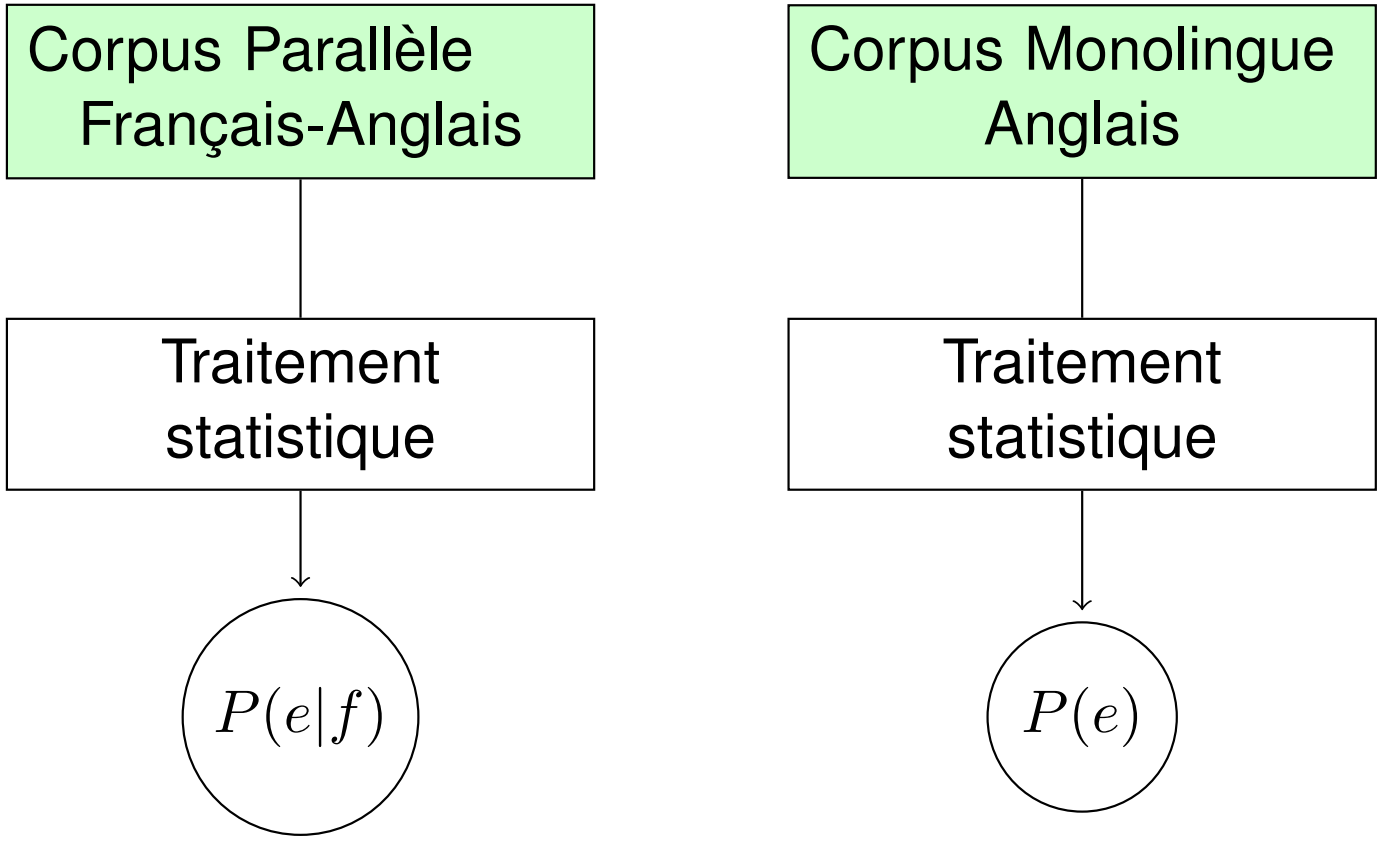
- Des aligneurs disponibles (Gale and Church, 1993 ; Moore, 2001)

Des ressources, des modèles, des algorithmes

Corpus Parallèle Français-Anglais

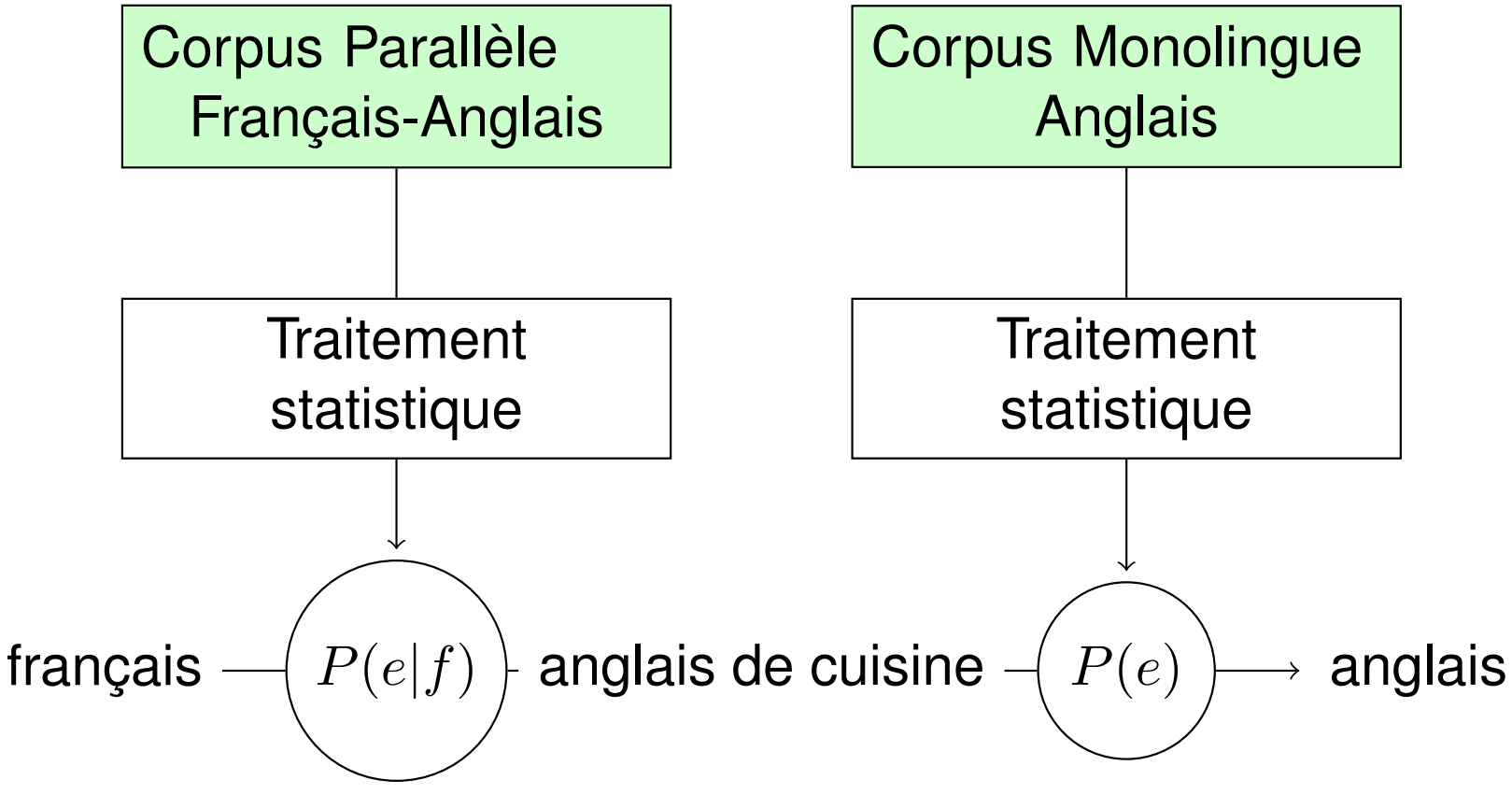
Corpus Monolingue Anglais

Des ressources, des modèles, des algorithmes

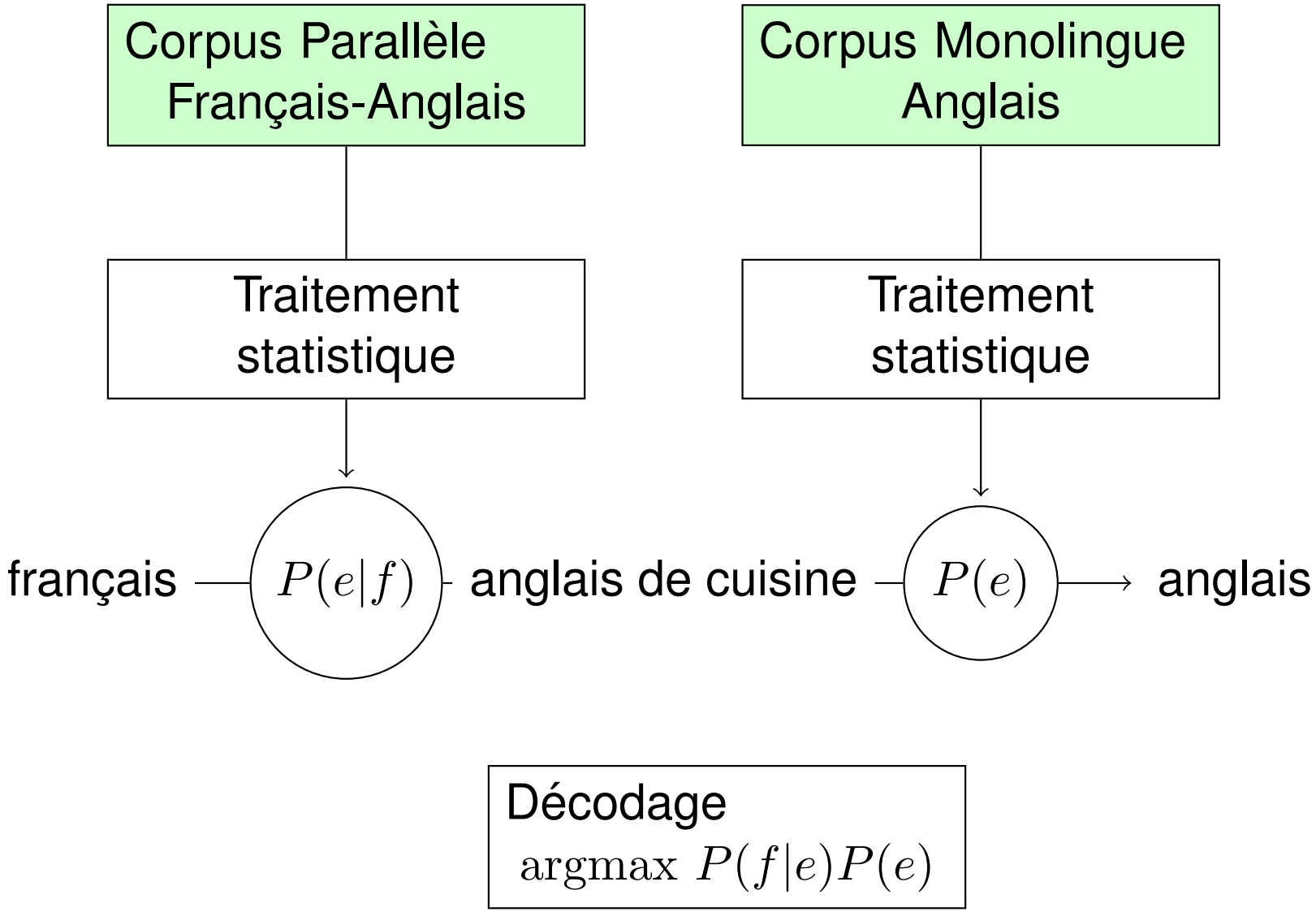


rali

Des ressources, des modèles, des algorithmes



Des ressources, des modèles, des algorithmes



Corpus alignés : le nerf de la guerre

- textes institutionnels :
  - débats parlementaires canadiens (anglais-français, anglais-inuktitut)
  - débats parlementaires européens (français, italien, espagnol, portugais, anglais, allemand, hollandais, danois, suédois, grec, finnois)
  - hong-kong (anglais-chinois)
  - santé-canada (anglais-français), Pan Health Organization (anglais-espagnol)
  - ...
- textes techniques
- *best sellers* :
  - Bible (2212), Coran (≥ 40), Catalogue IKEA (~ 30), Harry Potter (~ 30), ...
- internet

rali

Modèle de langue n-gramme

$$p(w = w_1, \ldots, w_N) \approx \prod_{i=1}^{N} p(w_i | w_{i-n+1}^{i-1})$$

Cas du modèle trigramme (n=2)

p(15 années de traduction en 15 minutes) = p(15) $\times$ p(années | 15) $\times$ p(de | 15 années) $\times$ p(traduction | années de) $\times$ p(en | de traduction) $\times$ p(15 | traduction en) $\times$ p(minutes | en 15)

Lire (Goodman, 2001), (Bengio et al., 2001)

Introduction des alignments *[?, ?]*

- estimation directe de $P(f|e)$?
- décomposition $P(f|e) = \prod_i P(f_i|e_i)$ trop simpliste
- décomposition via des alignements :

$$P(f|e) = \sum_a P(a, f|e)$$

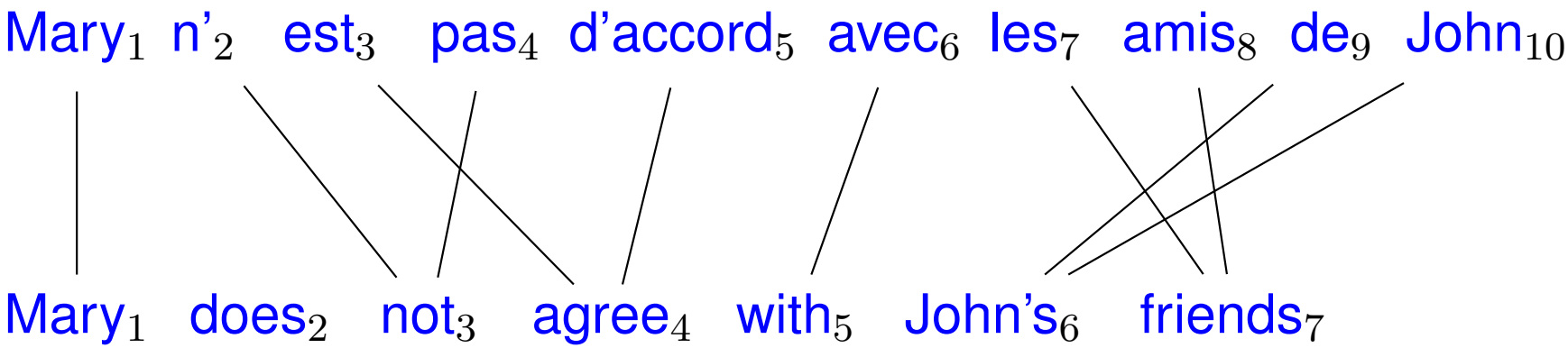
où a est un alignement entre e et f

Alignement de mots

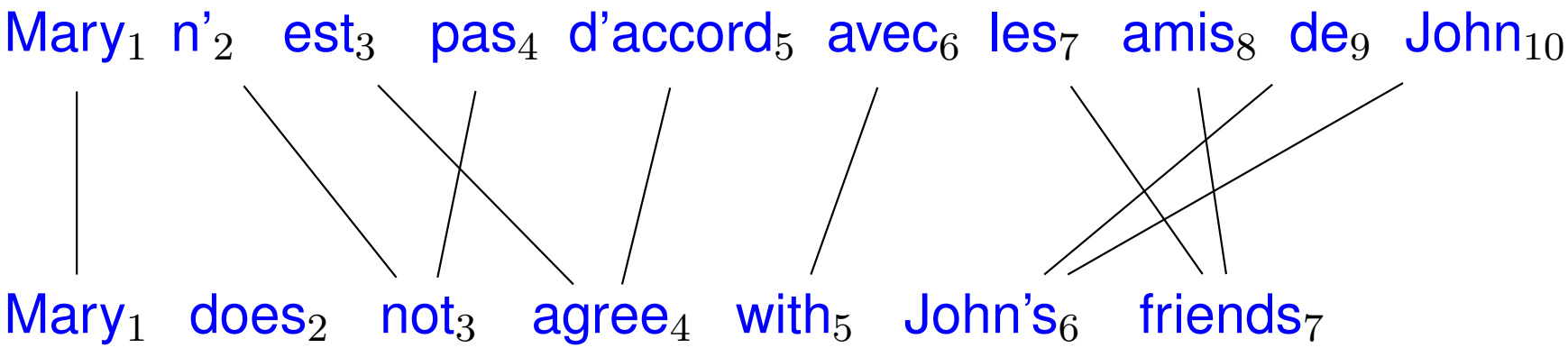
Mary$_1$ n'$_2$ est$_3$ pas$_4$ d'accord$_5$ avec$_6$ les$_7$ amis$_8$ de$_9$ John$_{10}$

Mary$_1$ does$_2$ not$_3$ agree$_4$ with$_5$ John's$_6$ friends$_7$

Alignement de mots



Alignement de mots



- un alignement $=$ relation sur $I \times J$.

$$a = \{(1,1),(2,3),(3,4),(4,3),(5,4)\ldots\}$$

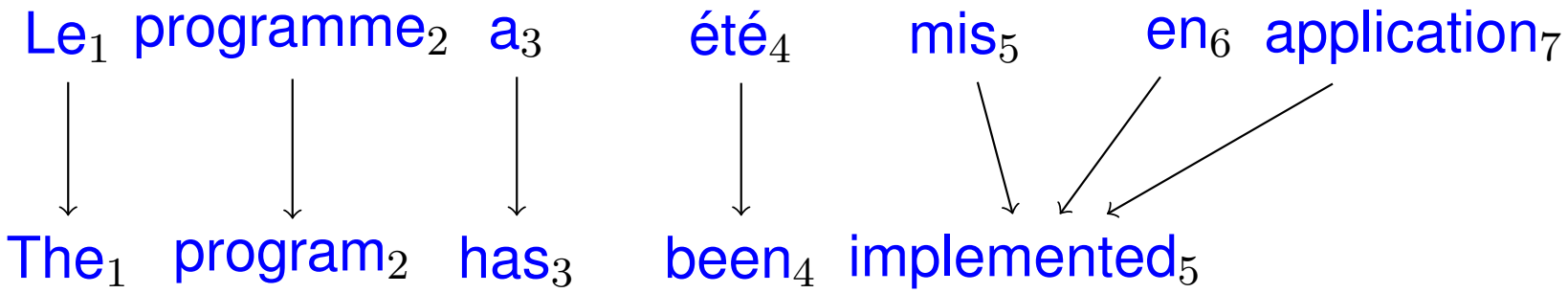
$2^{I \times J}$ relations possibles

- un alignement $=$ application partielle de I vers J :

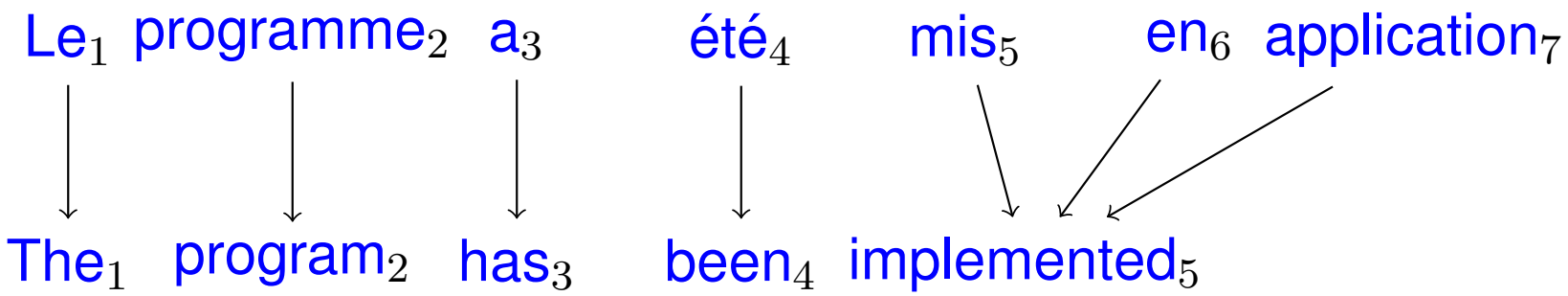
$$a = [1,3,4,3,4,5,7,7,6,6]$$

$\ll$ seulement $\gg$ I^{J+1} applications possibles

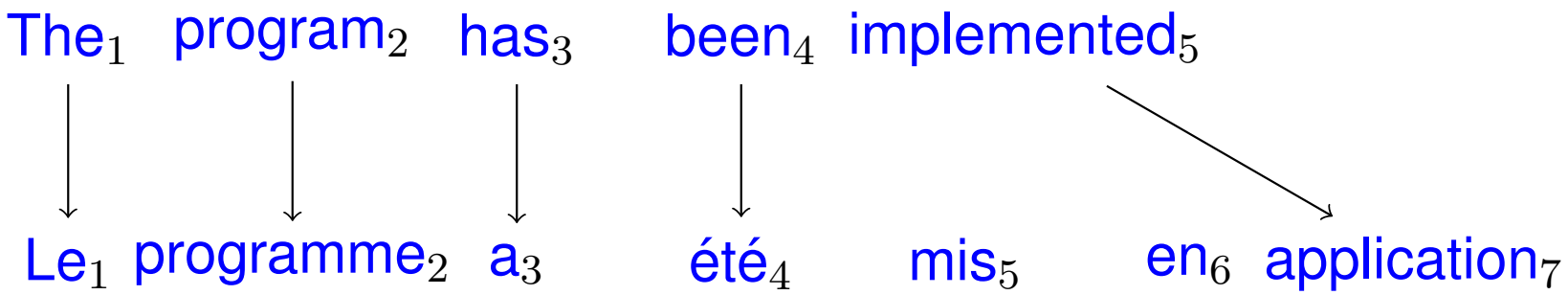
Problèmes des alignements de mots



Problèmes des alignements de mots

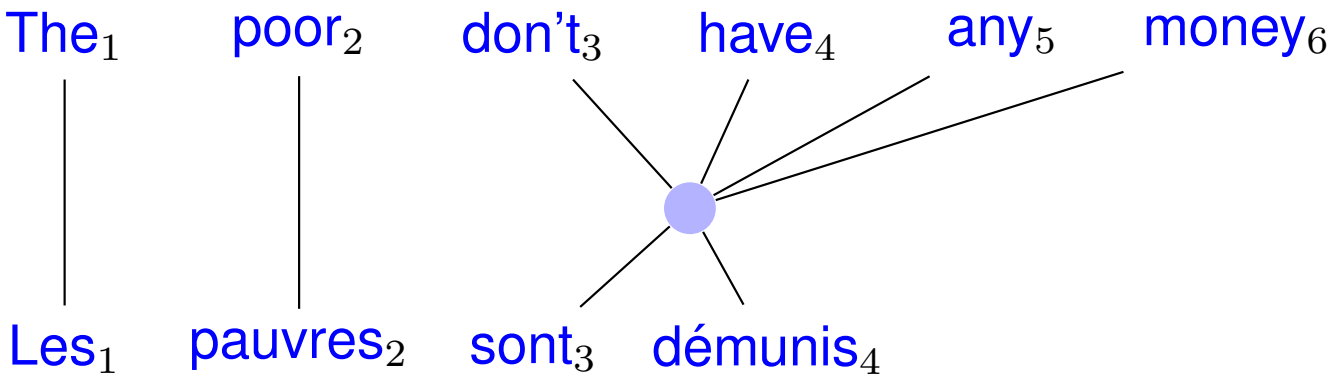


Mais :



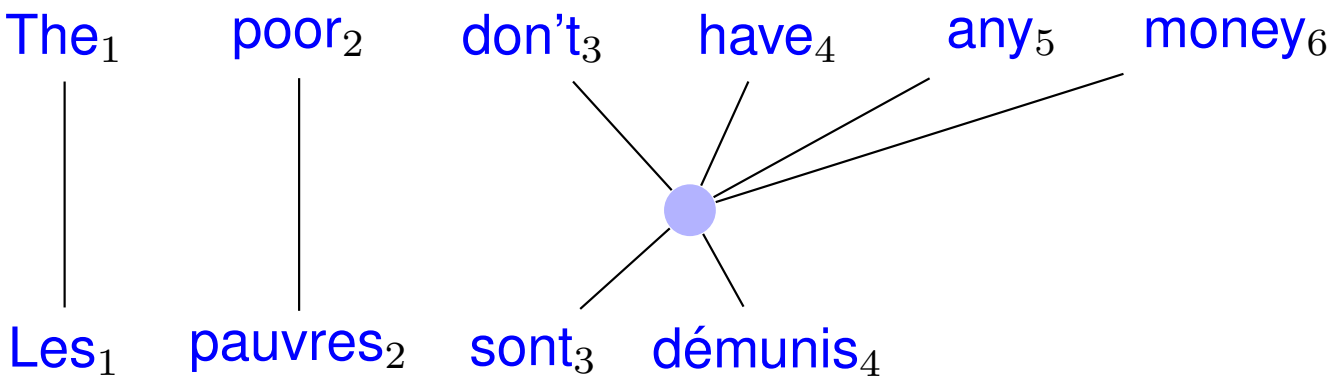
Modèles d'alignement non symétriques

Problèmes des alignements de mots

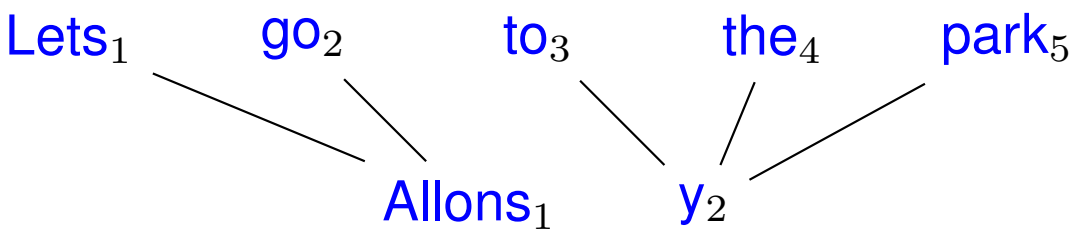


- Les alignements “à la IBM” ne sont pas toujours possibles

Problèmes des alignements de mots



- Les alignements “à la IBM” ne sont pas toujours possibles



- Présence de bruit dans les corpus

Modélisation avec alignements cachés

Notations

- $f_1^J = f_1 \dots f_J$ la phrase source (J mots)
- $e_1^I = e_1 \dots e_I$ la phrase cible (I mots)
- problème : décomposer $P(a, f|e)$

Modélisation avec alignements cachés

Notations

- $f_1^J = f_1 \dots f_J$ la phrase source (J mots)
- $e_1^I = e_1 \dots e_I$ la phrase cible (I mots)
- problème : décomposer $P(a, f|e)$

Structure du modèle génératif (IBM1,2 & HMM)

- choisir J sachant e_1^I
- pour chaque position $j \in [1 : J]$
  - choisir a_j sachant $J, a_1^{j-1}, f_1^{j-1}, e_1^I$
  - choisir f_j sachant $J, a_1^j, f_1^{j-1}, e_1^I$

$$P(a_1^J, f_1^J|e_1^I) = P(J|e_1^I) \prod_j P(a_j|a_1^{j-1}, f_1^{j-1}, e_1^I) P(f_j|a_1^j, f_1^{j-1}, e_1^I)$$

Processus Génératif

- NULL
- vendredi
- ,
- c’
- est
- badminton

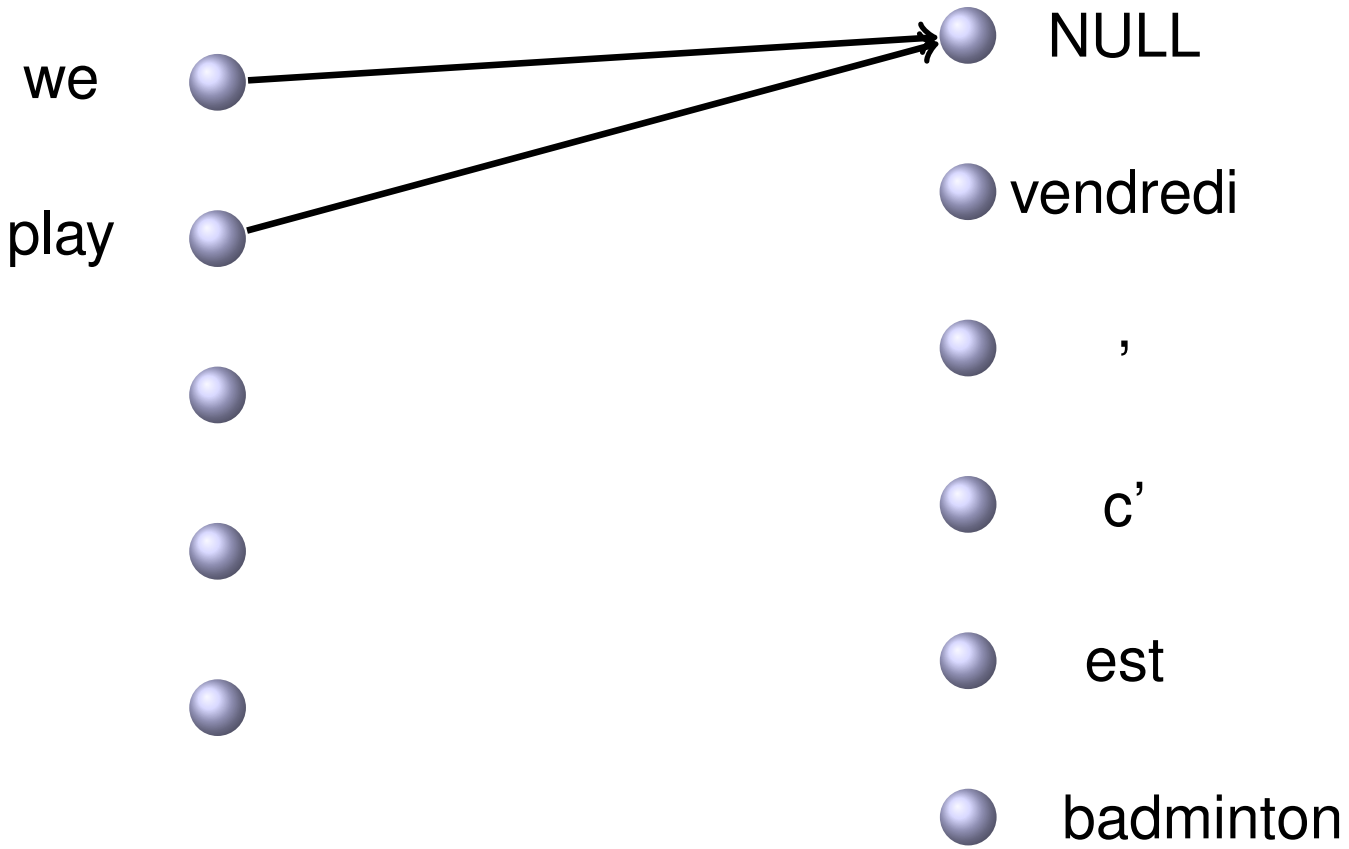
$$P(f, a|e) = P(J|I) \prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, I) P(f_j|a_1^{j}, f_1^{j-1}, J, I)$$

Processus Génératif

- NULL
- vendredi
- ,
- c’
- est
- badminton

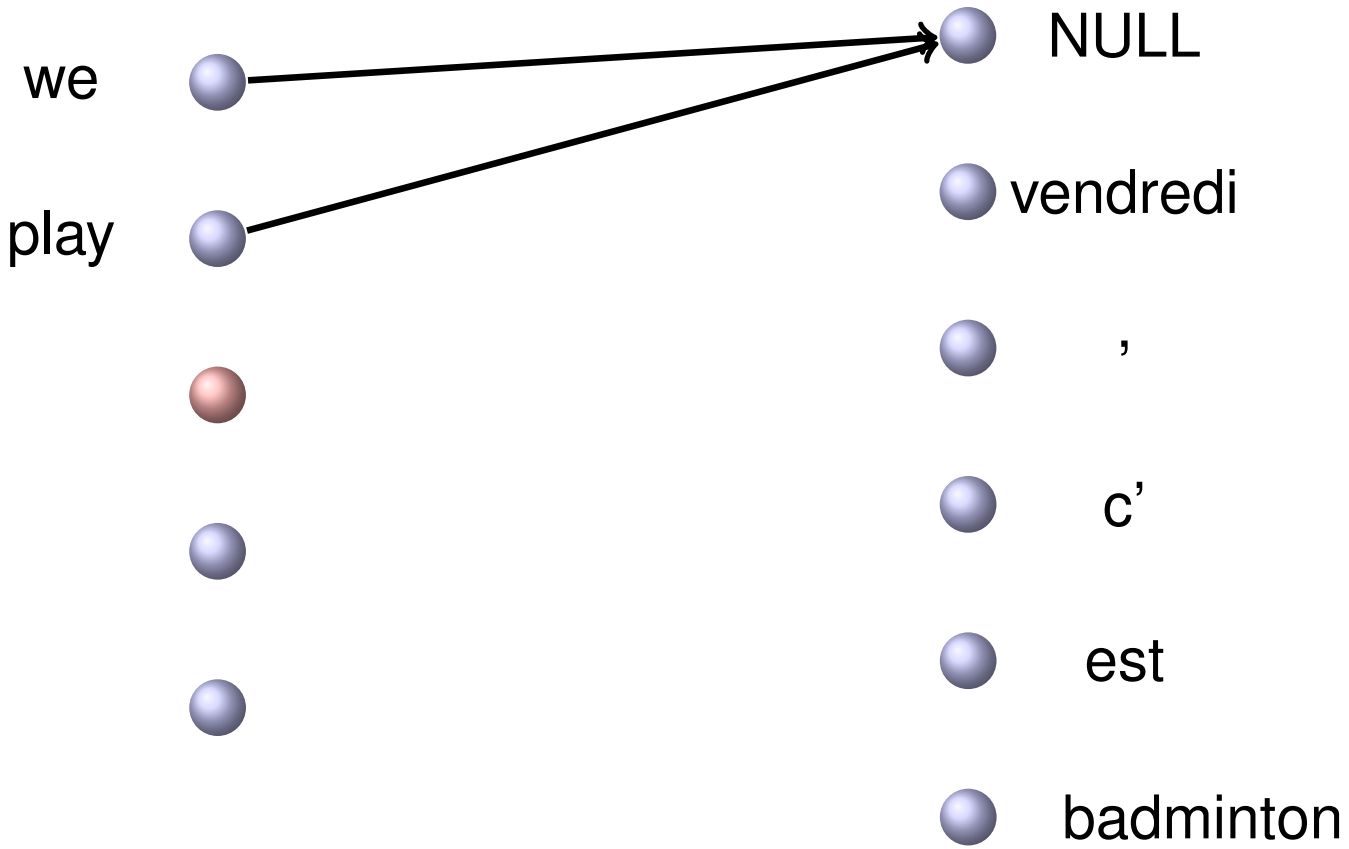
$$P(f, a|e) = P(J|I) \prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, I) P(f_j|a_1^{j}, f_1^{j-1}, J, I)$$

Processus Génératif



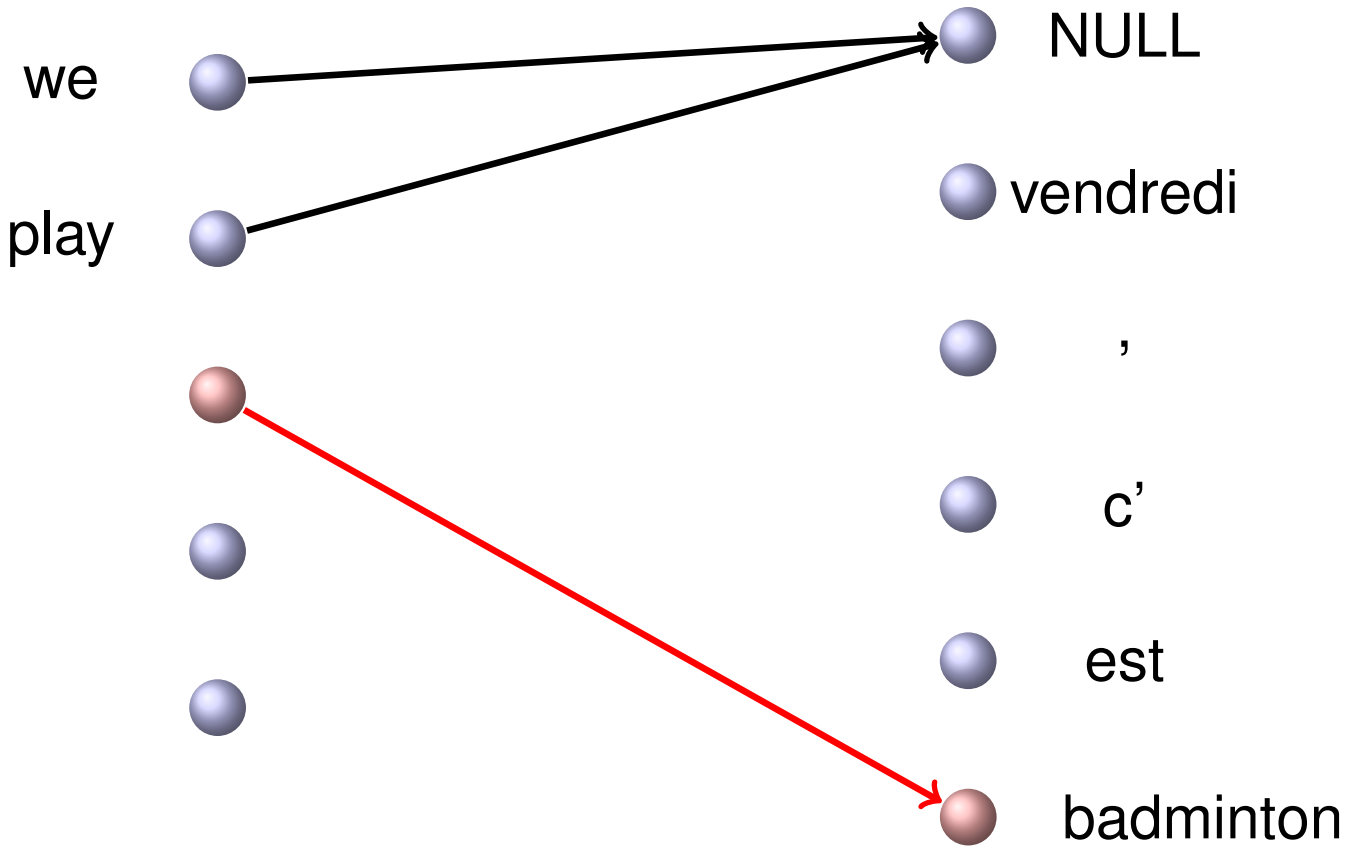
$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} P(a_j|a_1^{j-1},f_1^{j-1},J,I)P(f_j|a_1^{j},f_1^{j-1},J,I)$$

Processus Génératif



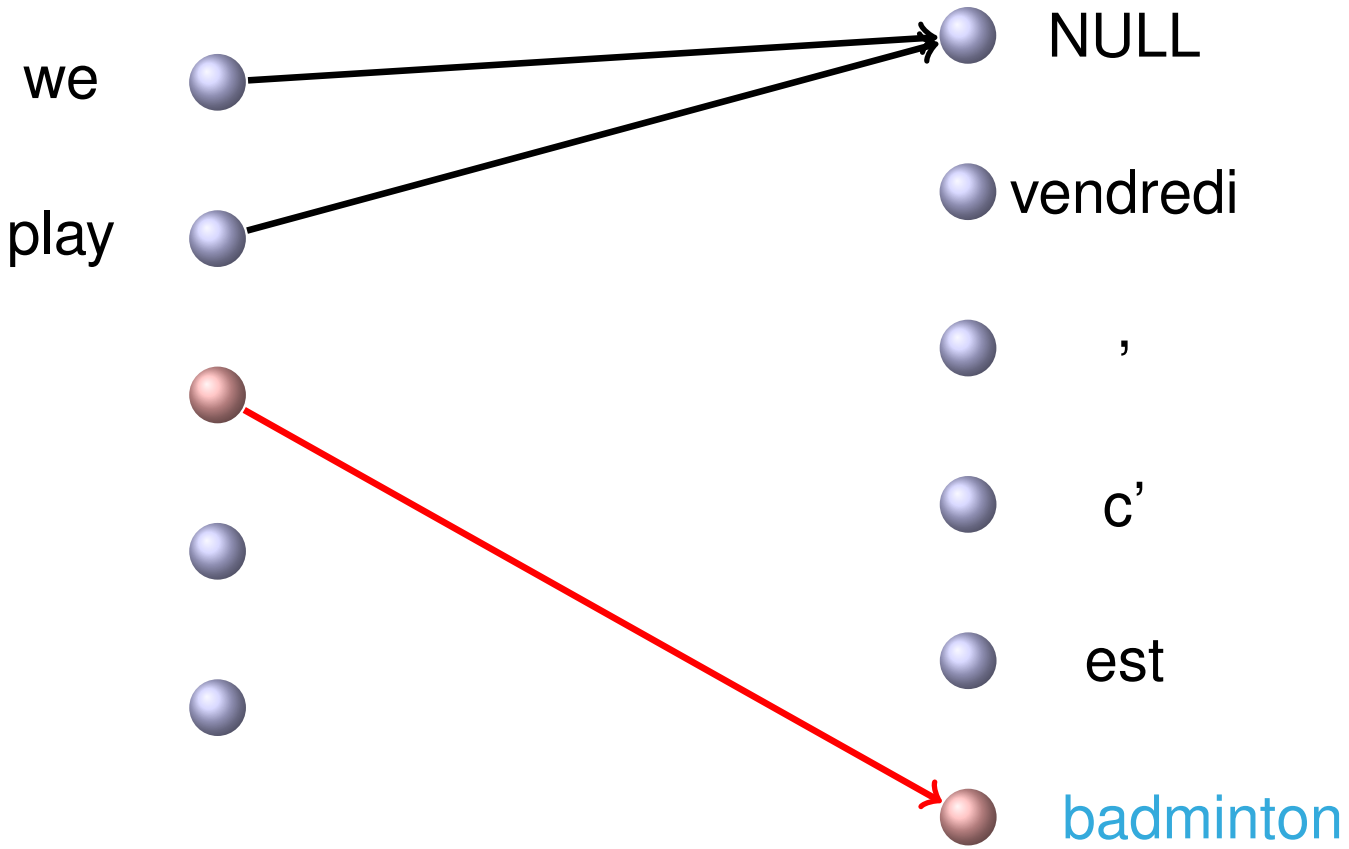
$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} P(a_j|a_1^{j-1},f_1^{j-1},J,I)P(f_j|a_1^{j},f_1^{j-1},J,I)$$

Processus Génératif



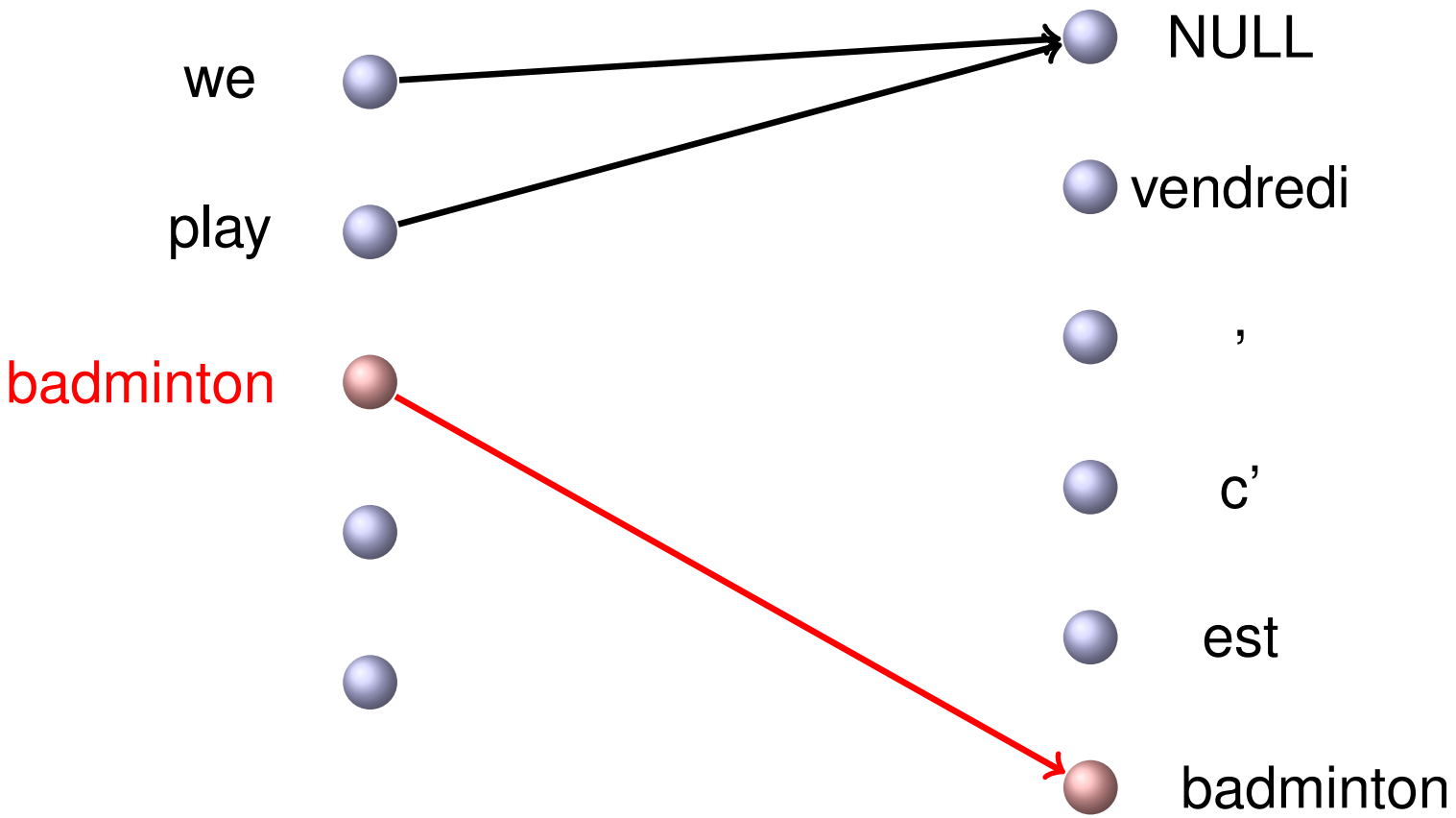
$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, I)P(f_j|a_1^{j}, f_1^{j-1}, J, I)$$

Processus Génératif



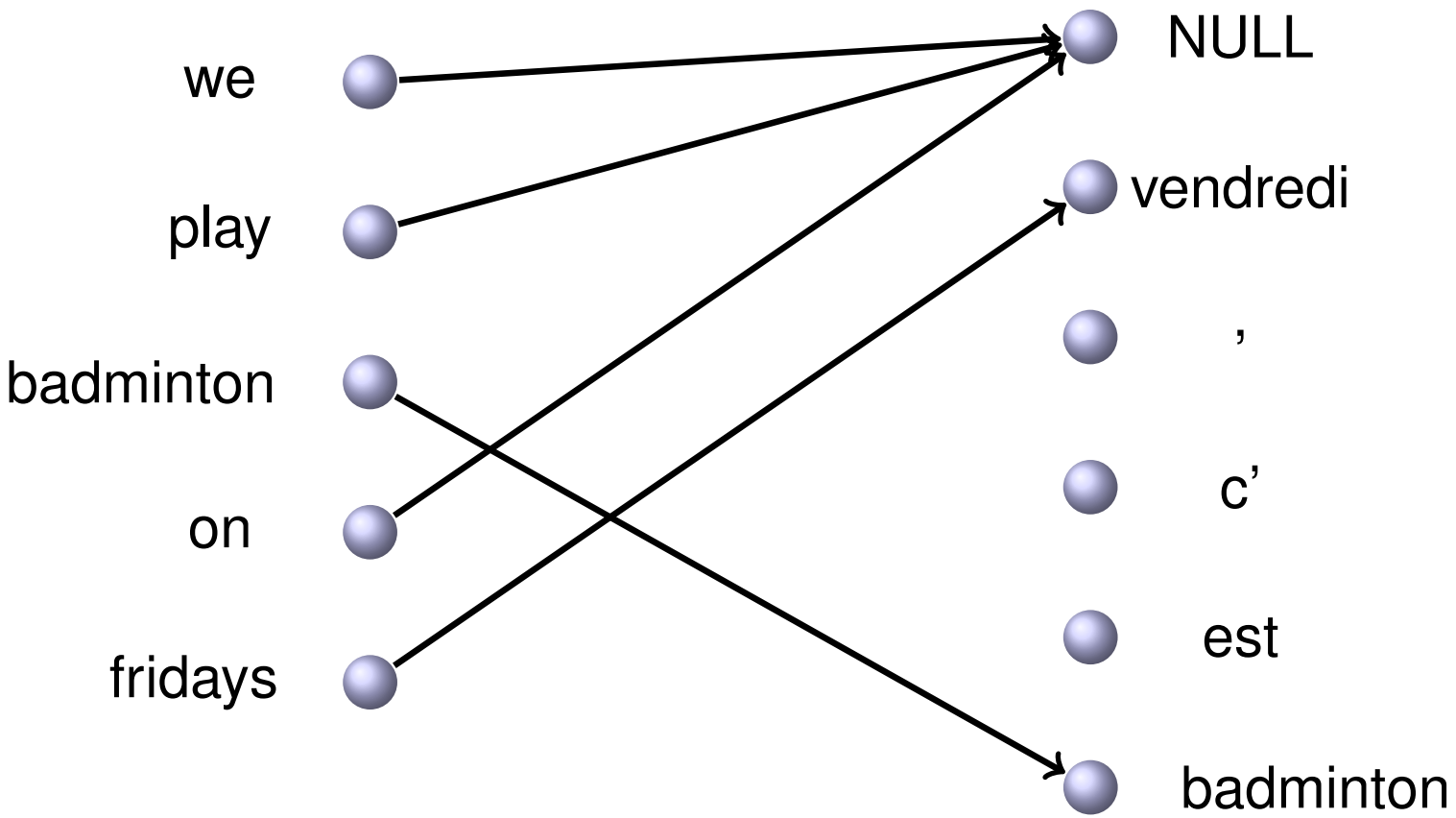
$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, I)P(f_j|a_1^{j}, f_1^{j-1}, J, I)$$

Processus Génératif



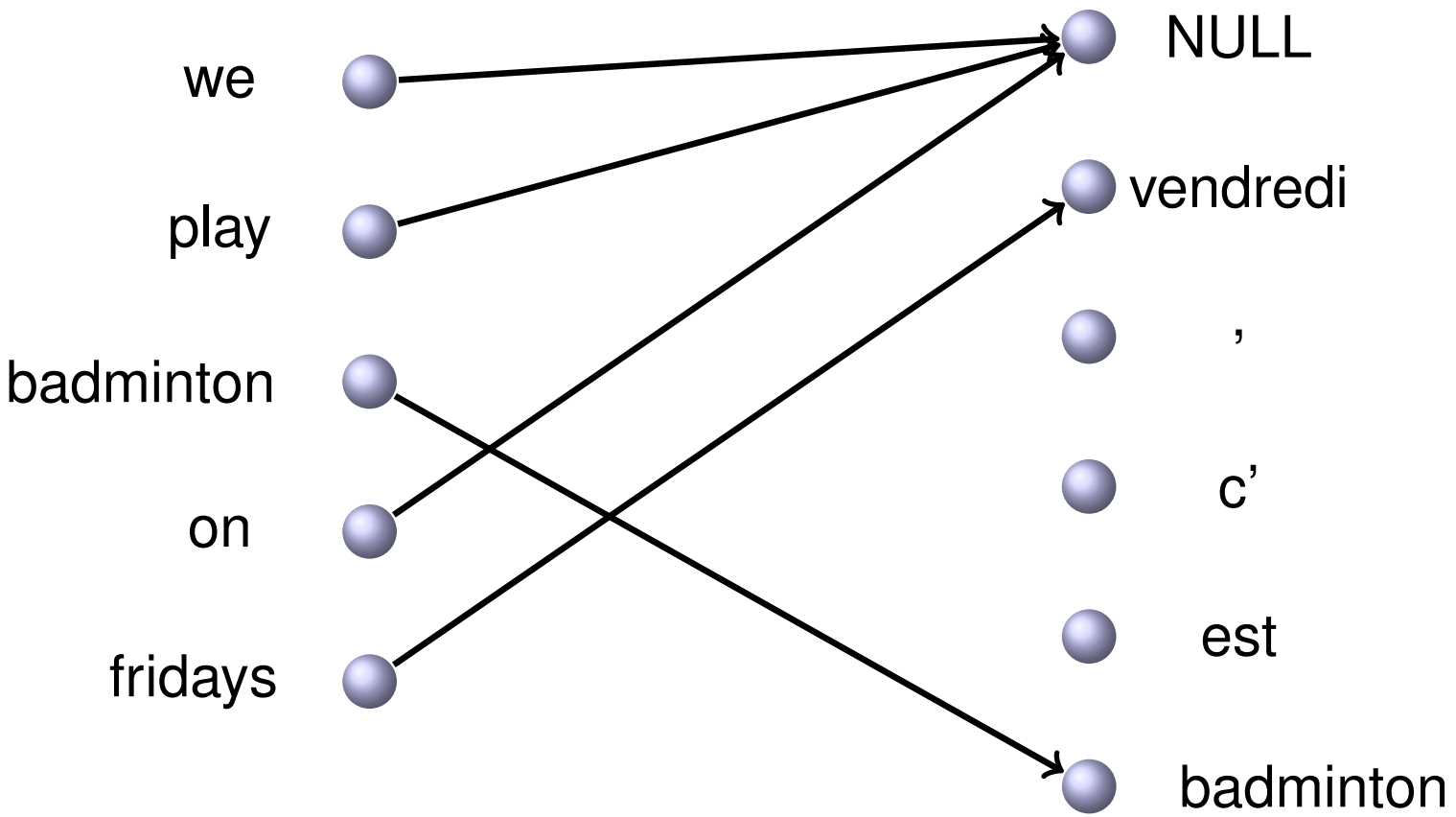
$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} P(a_j|a_1^{j-1},f_1^{j-1},J,I)P(f_j|a_1^{j},f_1^{j-1},J,I)$$

Processus Génératif



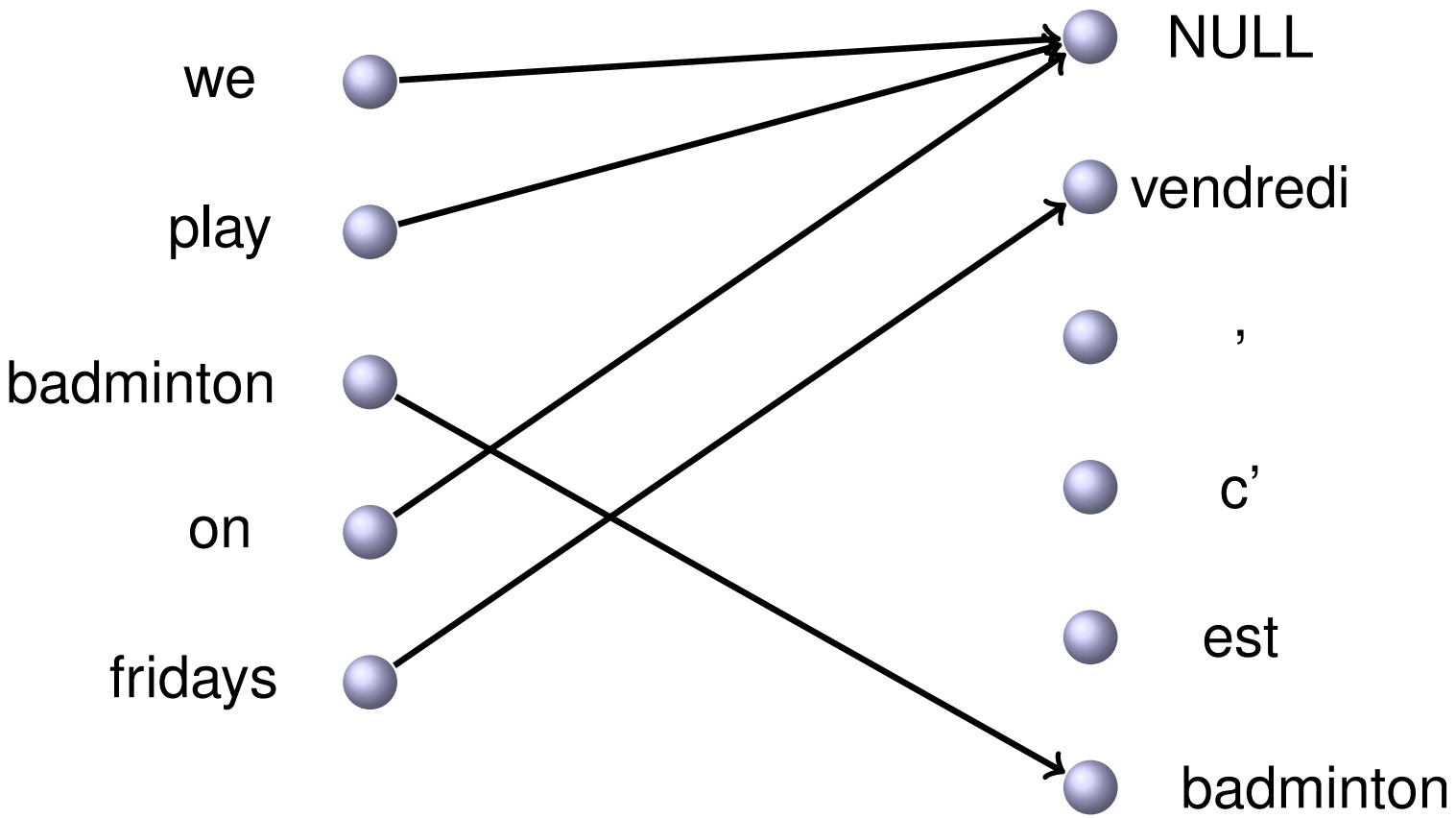
$$P(f, a|e) = P(J|I) \prod_{j=1}^{J} P(a_j|a_1^{j-1}, f_1^{j-1}, J, I) P(f_j|a_1^{j}, f_1^{j-1}, J, I)$$

Processus Génératif



$$P(f,a|e) = P(J|I)\prod_{j=1}^{J} p_a(a_j|a_{j-1},J)P(f_j|a_1^j,f_1^{j-1},J,I)$$

Processus Génératif



$$P(f, a|e) = P(J|I) \prod_{j=1}^{J} p_a(a_j|a_{j-1}, J) p_t(f_j|e_{a_j})$$

Deux finesses

Les mots “vides”

Traiter des mots source non alignables : ai et l’ dans :
j’ ai eu l’ occasion / I had occasion

- état fictif dans la cible (d’indice 0) atteint avec $P_0 = P(a_i = 0|a_{i-1}, J)$
- une distribution associée à cet état $P = P(f|\epsilon)$

Deux finesses

Les mots “vides”

Traiter des mots source non alignables : ai et l’ dans :
j’ ai eu l’ occasion / I had occasion

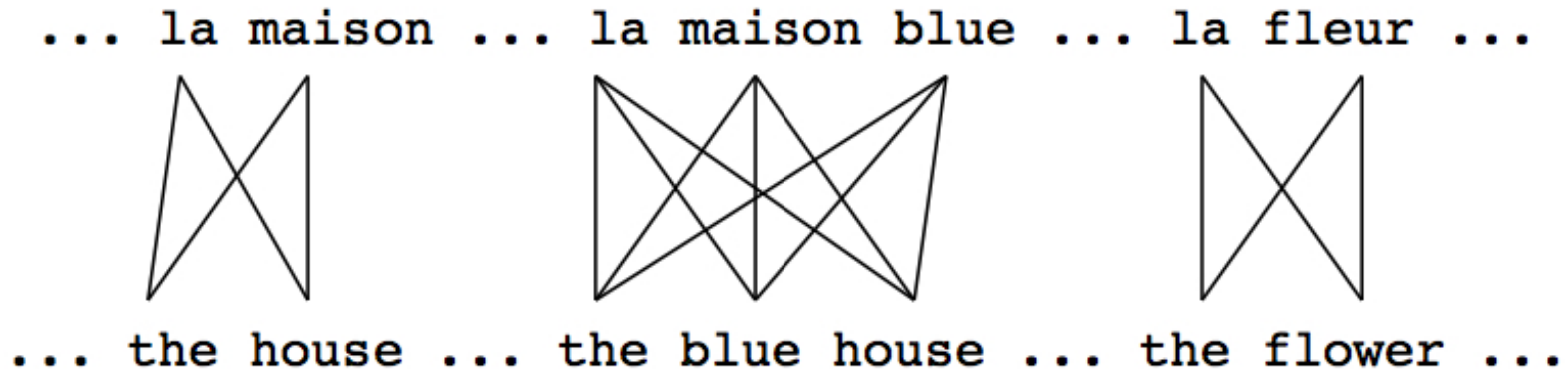
- état fictif dans la cible (d’indice 0) atteint avec $P_0 = P(a_i = 0|a_{i-1}, J)$
- une distribution associée à cet état $P = P(f|\epsilon)$

Modéliser les sauts

Rendre le modèle d’alignement indépendant des indices absolus : $\Rightarrow$ remplacer $P(a_i|a_{i-1})$ par $P(a_i - a_{i-1}|a_{i-1} - a_{i-2})$

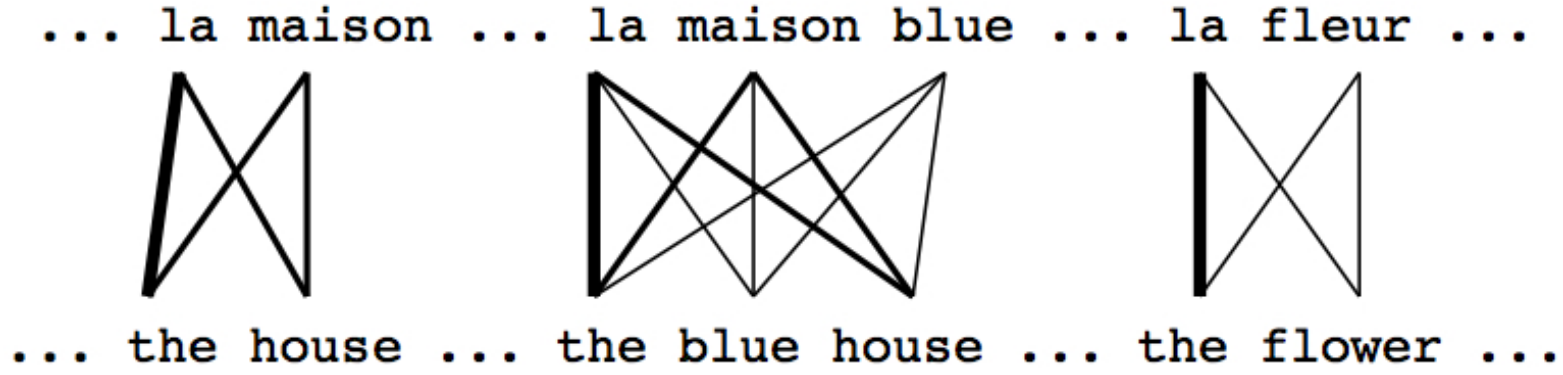
rali

Émergence des alignements



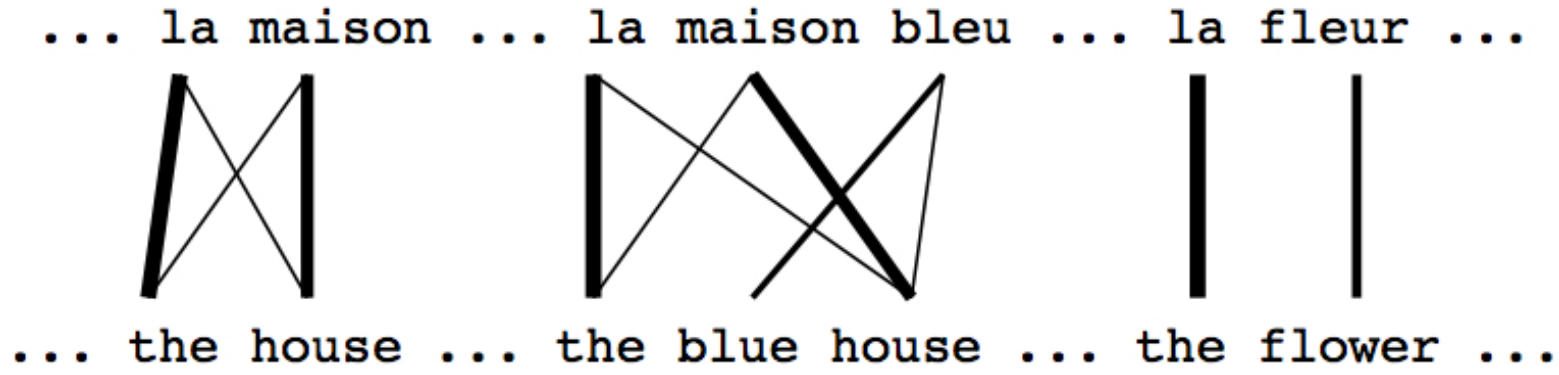
Tous les alignements sont également probables émergent...
se renforcent s'imposent (principe du « pigeonhole »)

Émergence des alignements



Tous les alignements sont également probables la/the, maison/house émergent... se renforcent s'imposent (principe du « pigeonhole »)

Émergence des alignements



Tous les alignements sont également probables émergent...
la/the, maison/house se renforcent s'imposent (principe du « pigeonhole »)

Émergence des alignements

Tous les alignements sont également probables émergent...
se renforcent bleue/blue, fleur/flower s'imposent (principe du ≪ pigeonhole ≫)

Estimation supervisée du modèle

- à alignements connus...
- ... les paramètres se déduisent par décompte :

$$\forall I \in [1 \dots I_{max}], J \in [1 \dots J_{max}], P(J|I) = \frac{n(I,J)}{n(I)}$$

$$\forall i, i' \in [1 \dots I_{max}], P(i'|i,J,I) = \frac{n(i,i')}{n(i)}$$

$$\forall e \in V_e, f \in V_f, P(f|e) = \frac{n(e,f)}{n(e)}$$

rali

Estimation par EM

Étape E(xpectation)

à paramètres connus (étape précédente) :

$$P(a_1^J|e_1^I, f_1^J) = \frac{P(a_1^J, f_1^J|e_1^I)}{\sum_a P(a_1^J, f_1^J|e_1^I)}$$

Le dénominateur se calcule par programmation dynamique.

Estimation par EM

Étape E(xpectation)

à paramètres connus (étape précédente) :

$$P(a_1^J|e_1^I, f_1^J) = \frac{P(a_1^J, f_1^J|e_1^I)}{\sum_a P(a_1^J, f_1^J|e_1^I)}$$

Le dénominateur se calcule par programmation dynamique.

Étape M(aximisation)

$$\forall I \in [1 \dots I_{max}], J \in [1 \dots J_{max}], P(J|I) = \frac{n(I, J)}{n(I)}$$

$$\forall i, i' \in [1 \dots I], P(i'|i, J, I) = \frac{\sum_k P(a^{(k)}|e^{(k)}, f^{(k)}) n^{(k)}(i, i')}{\sum_{i'} \sum_k P(a^{(k)}|e^{(k)}, f^{(k)}) n^{(k)}(i, i')}$$

$$\forall e, f, P(f|e) = \frac{\sum_{(k)} P(a^{(k)}|e^{(k)}, f^{(k)}) n^{(k)}(e, f)}{\sum_f \sum_{(k)} P(a^{(k)}|e^{(k)}, f^{(k)}) n^{(k)}(e, f)}$$

Les deux derniers termes se calculent par programmation dynamique (algorithme *Forward-Backward*)

rali

Distributions lexicales

the	(3/149)	(le,0.18) (la,0.15) (de,0.12)
minister	(2/27)	(ministre,0.8) (le,0.12)
people	(3/66)	(gens,0.25) (les,0.16) (personnes,0.1)
years	(3/24)	(ans,0.38) (années,0.31) (depuis,0.12)

$$\forall \mathrm{e}, \quad \sum_{\mathrm{f}} p(\mathrm{f}|\mathrm{e}) = 1$$

Calculer les alignements (à modèle connu)

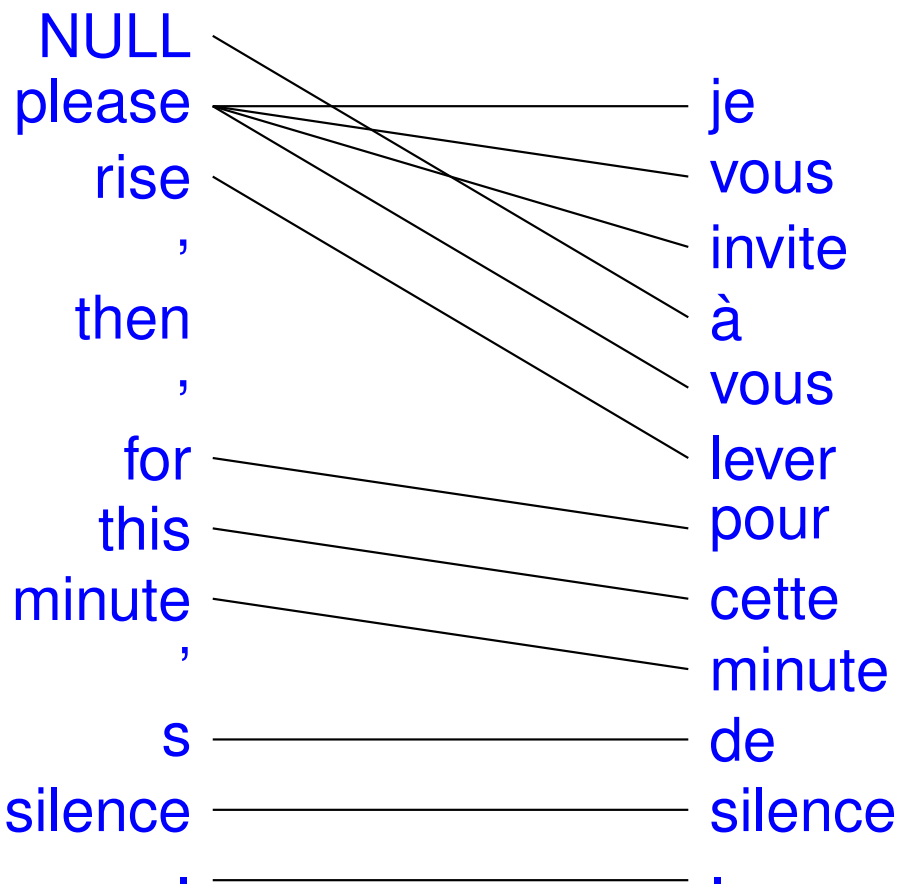
- $P(.|I)$ connu ; $P(.|a, I, J)$ connu ; $P(f|e)$ connu
- e_1^I et f_1^J sont observés
- trouver :

$$\begin{aligned} a^* &= \underset{a_1 \ldots a_J}{\operatorname{argmax}} P(a_1^J | f_1^J, e_1^I) \\ &= \underset{a_1 \ldots a_J}{\operatorname{argmax}} P(f_1^J, a_1^J | e_1^I) \\ &= \underset{a_1 \ldots a_J}{\operatorname{argmax}} P(J|I) \prod_j P(a_j|a_{j-1}) P(f_i|e_{a_j}) \end{aligned}$$

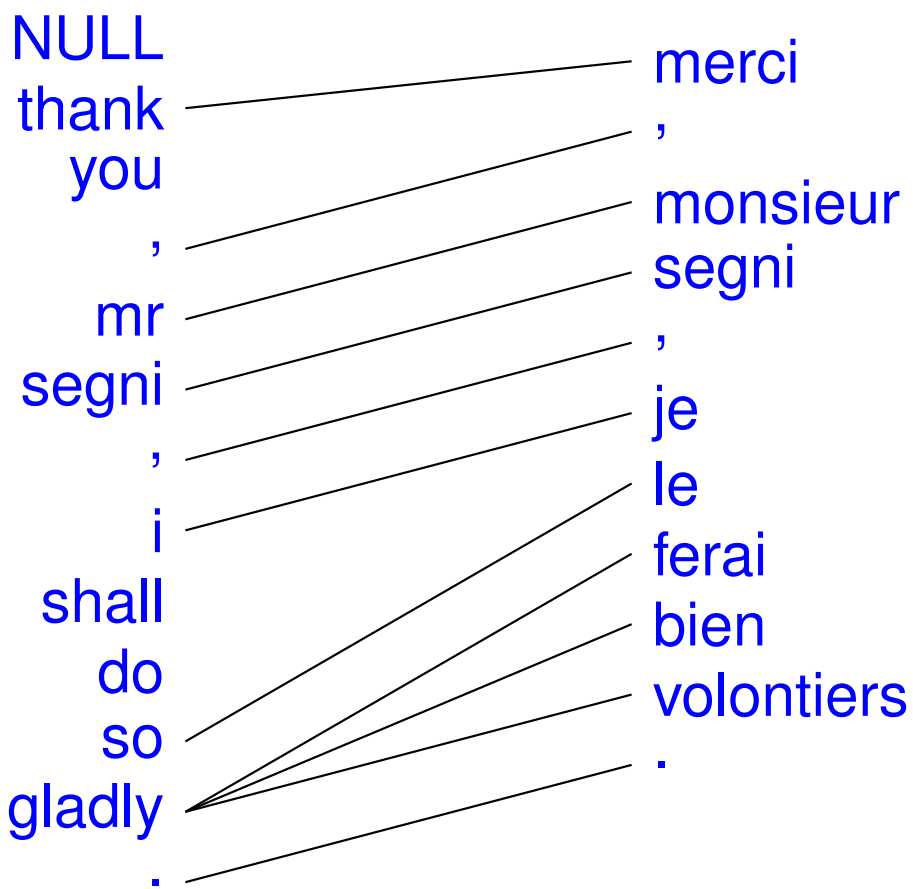
- Résolution par programmation dynamique (Viterbi)

$$\begin{cases} \delta(i,1) = P(a_1 = i), \forall i \in [1 \ldots I] \\ \delta(i,j) = max_{i' \in I} \delta(i', j-1) P(a_j = i | a_{j-1} = i') P(f_j|e_i) \forall i, j > 1 \end{cases}$$

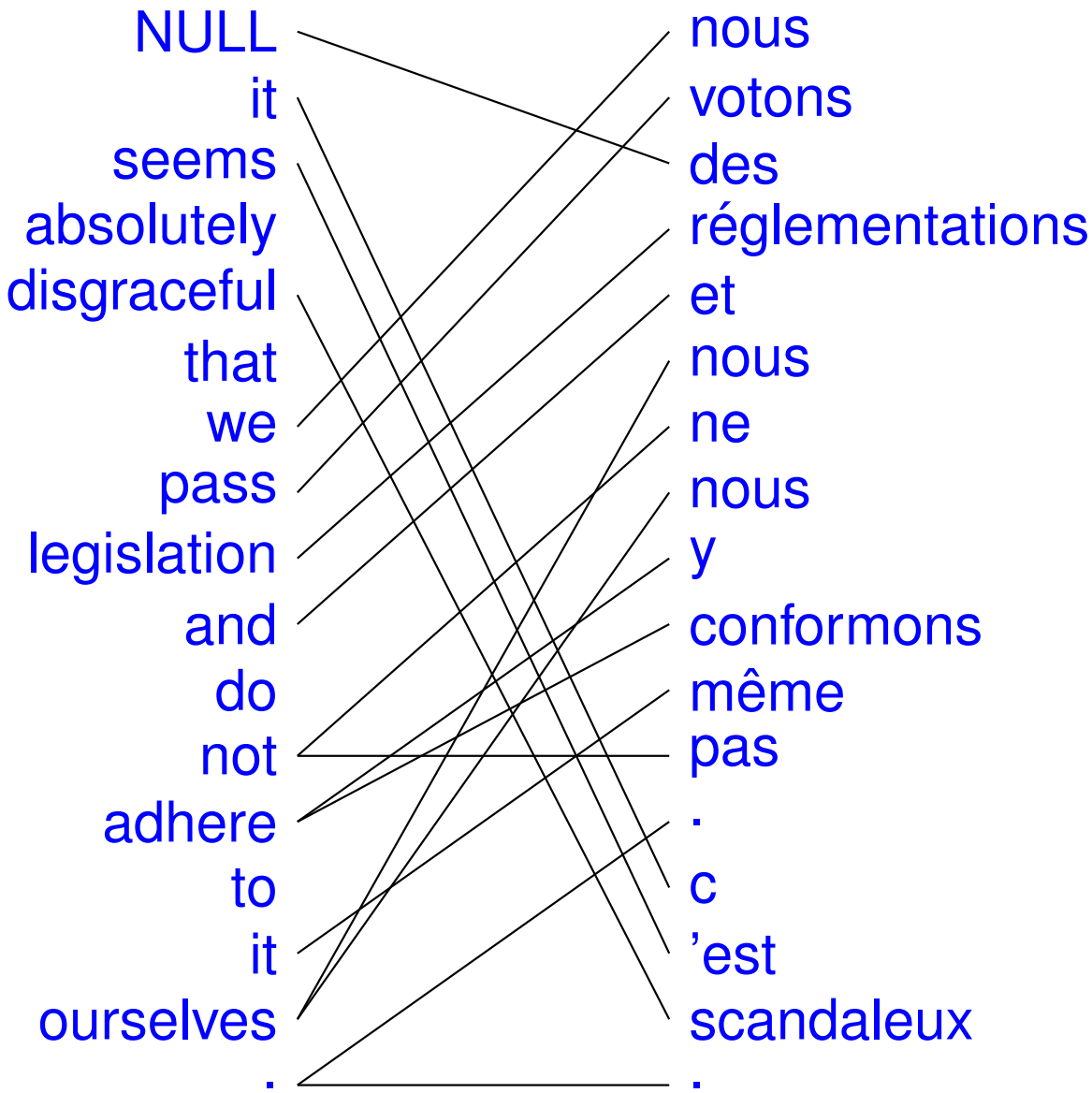
Des alignements... plus ou moins heureux



Des alignements... plus ou moins heureux



Des alignements... plus ou moins heureux



Pour en savoir plus...

- The mathematics of statistical machine translation (Brown & al, 1993) : publication de référence sur la traduction mot-à-mot et les modèles d'alignement
- A Statistical MT tutorial workbook ***[Knight, 1999b]*** : très pédagogique (disponible ici `http://www.statmt.org`)
- Giza, Giza++, Giza-pp : logiciel *open-source* pour la construction d'alignements
- `http://www.statmt.org`

rali

Vers les modèles de segment

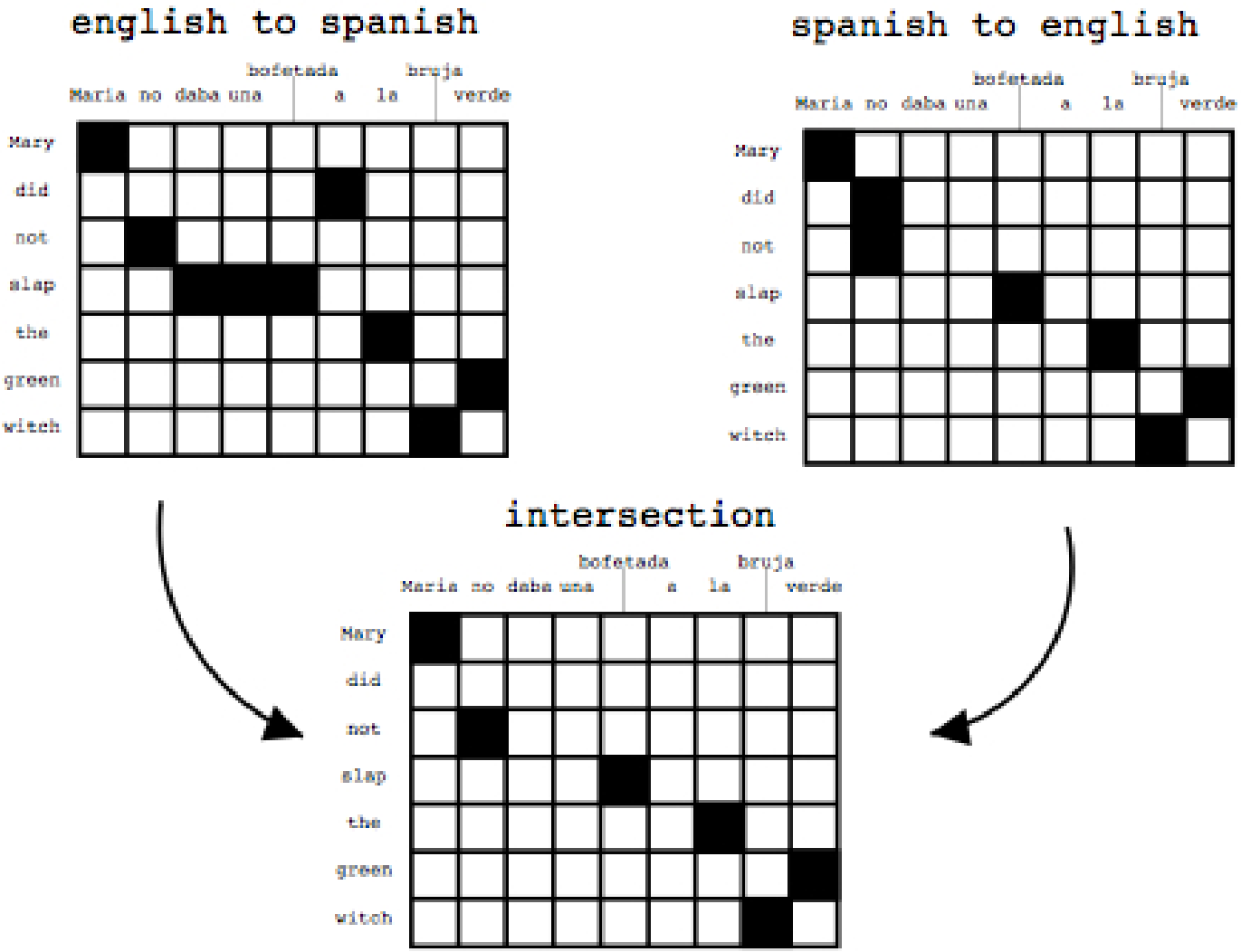
[Och and Ney, 2004, Koehn et al., 2003]

- Les alignements mot-à-mot sont problématiques
- Le modèle lexical $t(f|e)$ n'utilise pas de contexte :
  ex : Les poules du couvent couvent
- Prise en compte des formes figées (vue à l'entraînement)
  ex : . . . ont renoncé de guerre lasse à . . . — has finally given up trying
- Un modèle lexical $t(f_j|e_{i-2}e_{i-1}e_i)$ est trop complexe

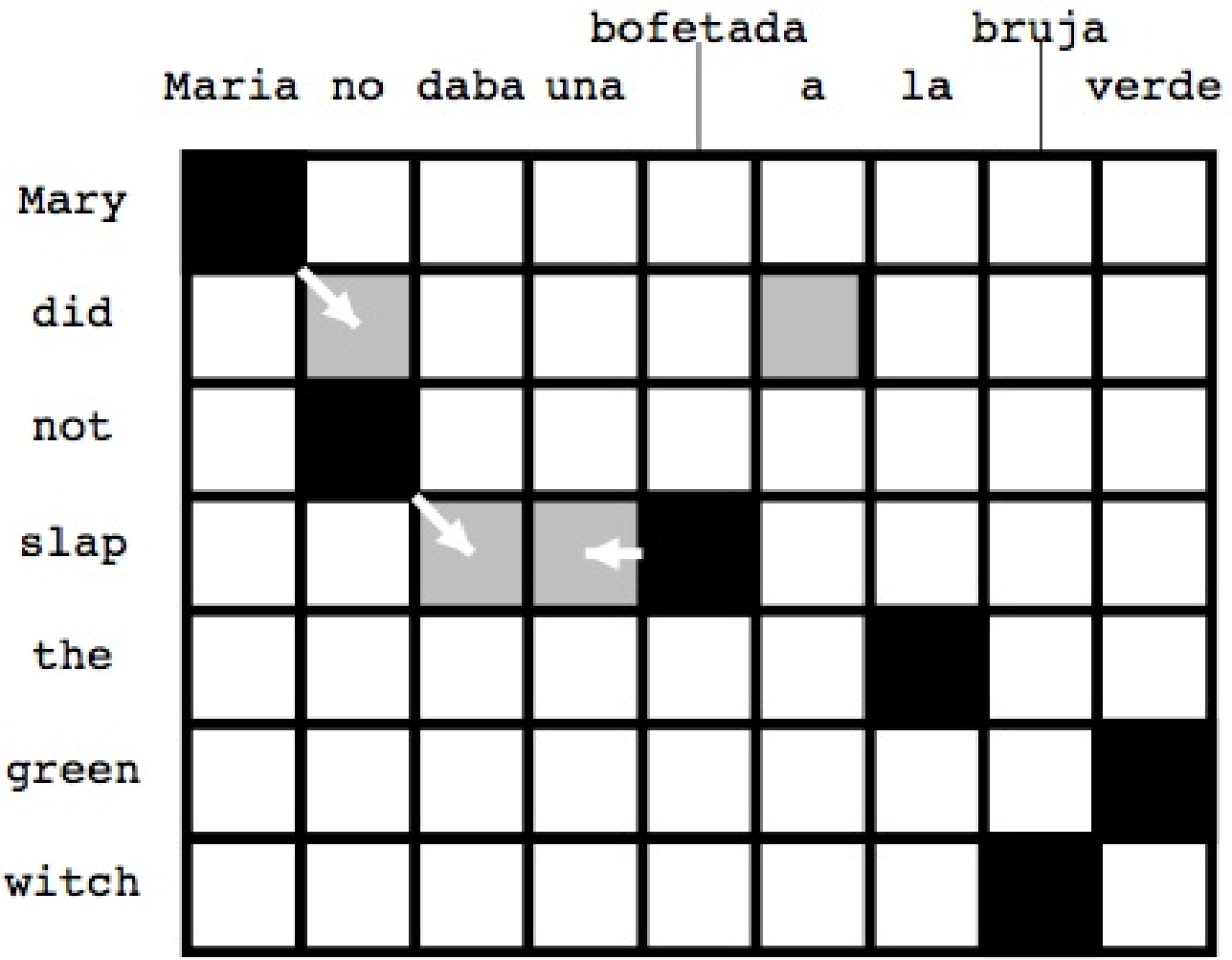
$\Rightarrow$ nouveau modèle de traduction, alignement de "blocs de mots" (segments).
: Apprentissage du modèle

  - acquisition des segments
  - modèle probabiliste à base de segments

Extraction de segments



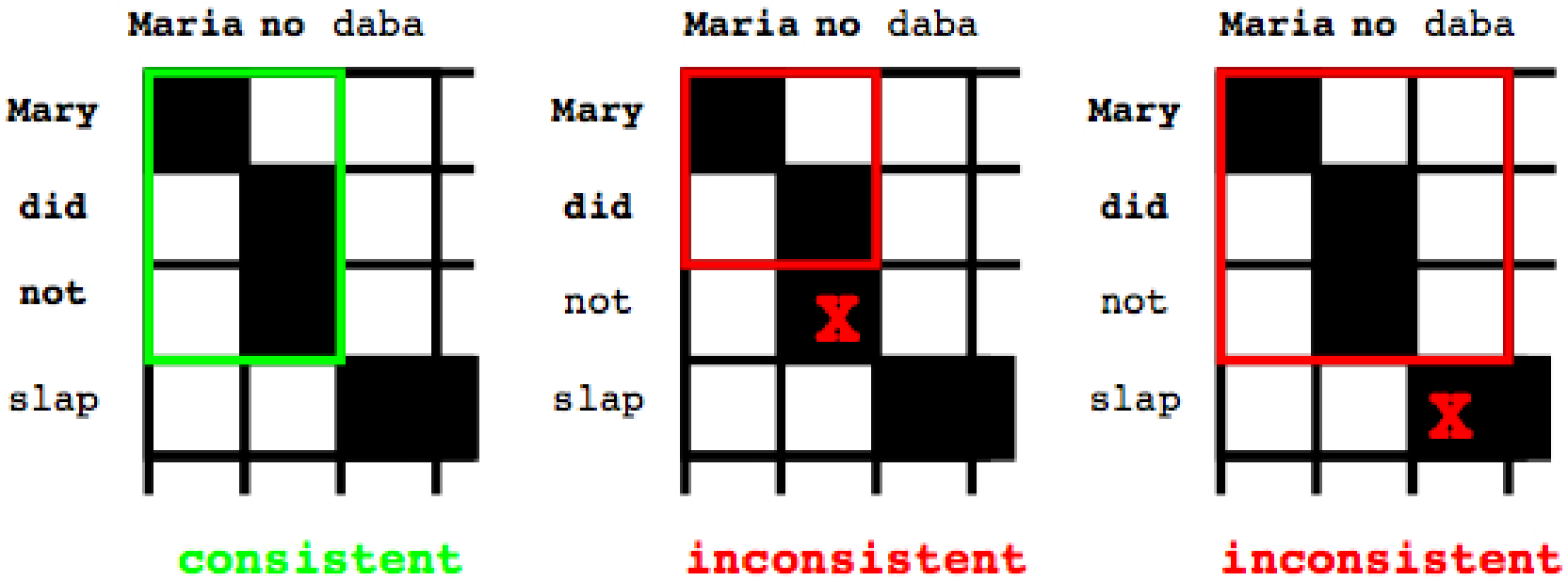
Extraction de segments



Extraction des segments

Les alignements symétrisés

Extraction des segments



Les contraintes de cohérence

$$\forall e_i \in \tilde{e}, (e_i, f_j) \in A \Rightarrow f_j \in \tilde{f}$$
$$\forall f_j \in \tilde{f}, (e_i, f_j) \in A \Rightarrow e_i \in \tilde{e}$$

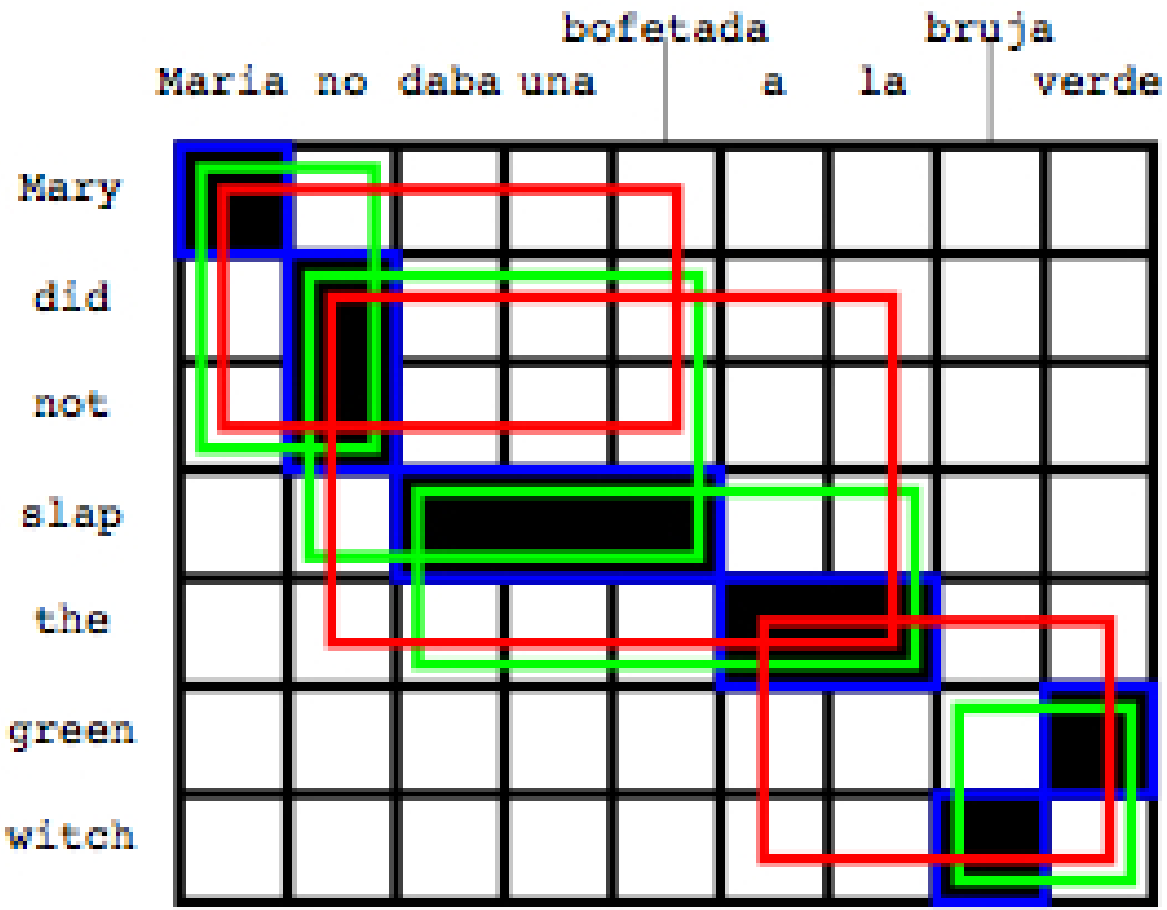
Extraction des segments

(Mary,Maria), (did not, no), (slap, daba una bofetada)...

Extraction des segments

(Mary did not,Maria no), (did not slap, no daba una bofetada)...

Extraction des segments



(Mary did not slap,Maria no daba una bofetada)...

Les scores d'un fragment

- Au maximum de vraisemblance :
  - $P(\tilde{f}|\tilde{e}) = \frac{c(\tilde{f},\tilde{e})}{c(\tilde{e})}$
  - $P(\tilde{e}|\tilde{f}) = \frac{c(\tilde{e},\tilde{f})}{c(\tilde{f})}$

  $\Rightarrow$ estimateurs très optimistes pour les longs segments
- Autres options :
  - $P(\tilde{f}|\tilde{e}) = P_{IBM}(\tilde{f}|\tilde{e})$
  - $P(\tilde{e}|\tilde{f}) = P_{IBM}(\tilde{e}|\tilde{f})$
- Pourquoi choisir ? $\Rightarrow$ combinaison des scores

rali

Combinaison des scores et tuning

- Nouveau modèle de traduction (indépendance entre segments) :

$$P(e|f) = \sum_{a=s1\ldots s_k} \prod_{i=1}^{k} P(\tilde{f}_i|\tilde{e}_i)$$

$$\approx \max_{a=s1\ldots s_k} \prod_{i=1}^{k} P(\tilde{f}_i|\tilde{e}_i)$$

- Modèles probabilistes individuellement imprécis

$\Rightarrow$ Pondération de leur influence :

$$e^* = \operatorname*{argmax}_e \prod_k P_k(f,e)^{\lambda_i}$$

$$e^* = \operatorname*{argmax}_e \sum_k \lambda_k \log P_k(f,e)$$

- Comment déterminer les coefficients λ_i ?

Optimisation du système : calcul des λ

- À la main ?
- Boucle exploratoire :
  1. Choisir λ_k initiaux
  2. Faire un décodage avec ces valeurs
  3. Obtenir une solution et calculer son score
  4. Modifier les λ_k et recommencer à l'étape 2
  5. Terminer si le score ne s'améliore plus

$\Rightarrow$ Algorithmes itératifs de recherche

rali

Les ingrédients d'un modèle de segment

- Modèle de traduction :

$P(\tilde{f}|\tilde{e})$ traduction segments $e \rightarrow f$
$P(f|e)$ traduction de mots $e \rightarrow f$ (modèle lexical type IBM1)
$P(\tilde{e}|\tilde{f})$ traduction segments $f \rightarrow e$
$P(e|f)$ traduction de mots $e \rightarrow f$ (modèle lexical type IBM1)
e constante $\rightarrow$ pénalité sur le nombre de segments

  - + modèles de distortions (une autre fois)

- Modèle de langage : $P(e)$
- constante 1 $\rightarrow$ pénalité de longueur

rali

La table des segments

Scores : $P(\tilde{f}|\tilde{e})$, $P(e|f)$, $P(\tilde{e}|\tilde{f})$, $P(f|e)$ et e

quelques traductions de “A big”

```
A big ||| Le grand ||| 0.0106383 0.000152962 0.166667 0.00405915 2.718
A big ||| Un des principaux ||| 0.0434783 0.0005689 0.166667 1.56536e-05 2.718
A big ||| Un grand ||| 0.00961538 0.00957428 0.166667 0.0300893 2.718
A big ||| Une grande ||| 0.0108696 0.00360665 0.166667 0.0208976 2.718
A big ||| ont une grande ||| 0.0217391 1.12938e-05 0.166667 3.79597e-06 2.718
A big ||| une grande ||| 0.000256345 1.12938e-05 0.166667 0.00211983 2.718
```

rali

La table des segments (suite)

467 traductions de “European Commission”

```
European Commission ||| Commission européenne ||| 0.752696 0.812097 0.749849 0.455413 2.718
European Commission ||| Commission ||| 0.00265859 0.00194196 0.0511501 0.952132 2.718
European Commission ||| la Commission européenne ||| 0.0426116 0.812097 0.0352603 0.0174883 2
European Commission ||| Commission européenne , ||| 0.17041 0.812097 0.0195218 0.0364258 2.71
European Commission ||| de la Commission européenne ||| 0.0625 0.812097 0.0160412 0.00229579 2
```

38 traductions inverses de “Commission européenne”

```
European Commission ||| Commission européenne ||| 0.752696 0.812097 0.749849 0.455413 2.718
Commission ||| Commission européenne ||| 0.116208 0.490344 0.00548883 0.00587199 2.718
the European Commission ||| Commission européenne ||| 0.0095701 0.0437849 0.0119704 0.455413 2
Commission ’s ||| Commission européenne ||| 0.00592435 0.00389219 0.0137227 0.00378834 2.718
Commission is ||| Commission européenne ||| 0.00303813 0.000335368 0.0036914 4.97013e-05 2.71
```

La table des segments (suite et fin)

672 traductions de ’ !’ ! ! !

```
! ||| ! ! ! ||| 0.375 0.588351 0.000338181 0.462852 2.718
! ||| ! ! ||| 0.153846 0.588351 0.000225454 0.598358 2.718
! ||| ! ||| 0.534388 0.588351 0.731372 0.773536 2.718
! ||| : non ! ||| 0.5 0.588351 0.000112727 2.60435e-07 2.718
...
! ||| , dit-on partout ! ||| 1 0.588351 0.000112727 4.76404e-12 2.718
! ||| , exigez que ||| 0.5 5.69e-05 0.000112727 1.92463e-10 2.718
! ||| , exigez ||| 0.333333 5.69e-05 0.000112727 1.20609e-08 2.718
! ||| , il est primordial que la ||| 0.333333 5.69e-05 0.000112727 3.20037e-15 2.718
! ||| , il est primordial que ||| 0.0277778 5.69e-05 0.000112727 8.33407e-14 2.718
...
! ||| Messieurs , il est primordial que la ||| 1 5.69e-05 0.000112727 4.92856e-19 2.718
! ||| Messieurs , il est primordial ||| 1 5.69e-05 0.000112727 8.04285e-16 2.718
...
```

Note : 1 million de paires de phrases $\sim$ 40 millions de paramètres . . .

Decoding

- Monotonous decoding :
  - efficient
  - no reordering allowed
- Decoding with distortion
  - argmax NP difficult (even with IBM1 !)
  - heuristic methods (A^* etc)
- Search space too big
  - filtering unlikely hypotheses

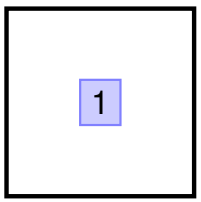
rali

This beautiful plant is unique

transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante		
plante magnifique		

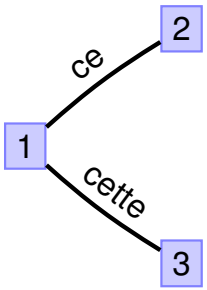
language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	

1

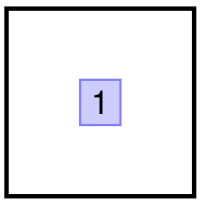


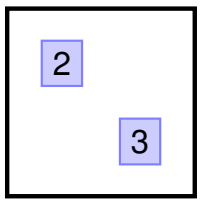
This beautiful plant is unique

transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

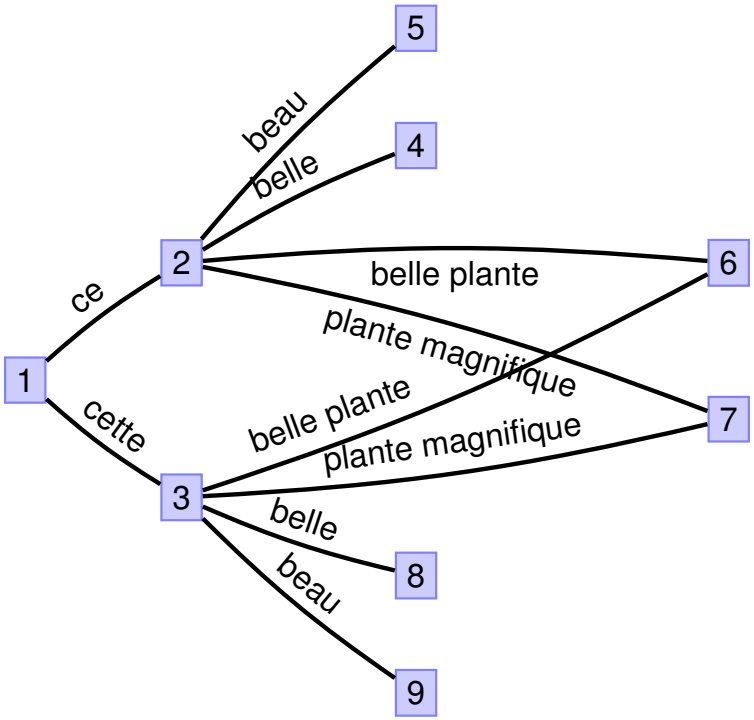


language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	



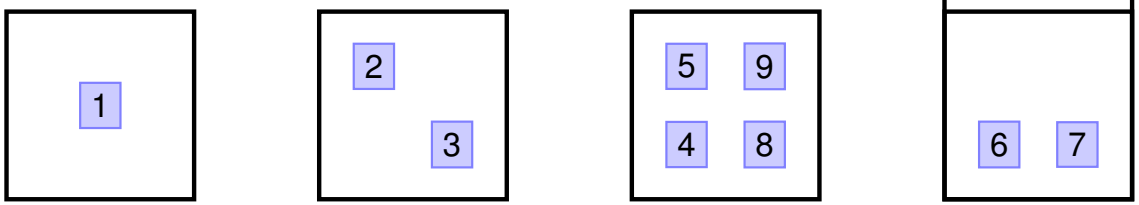


This beautiful plant is unique

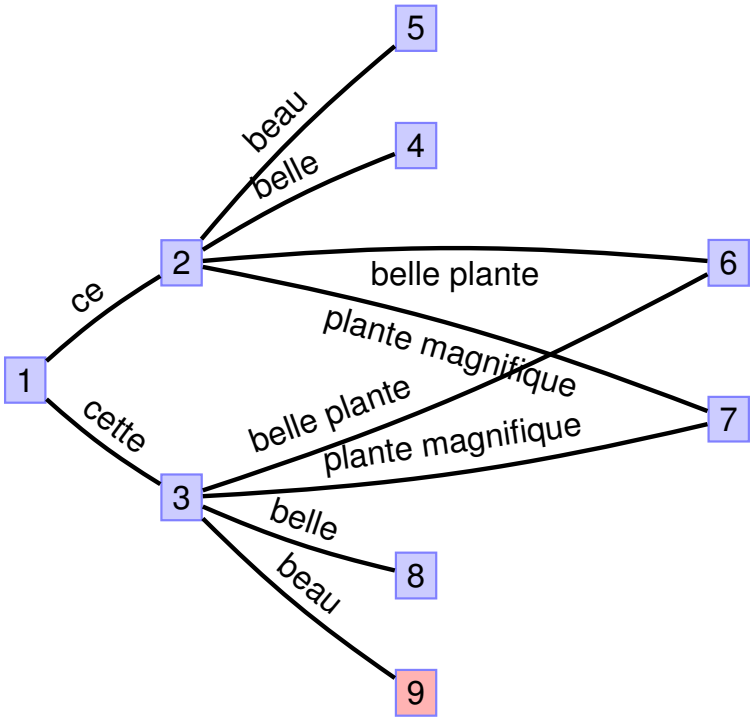


transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	

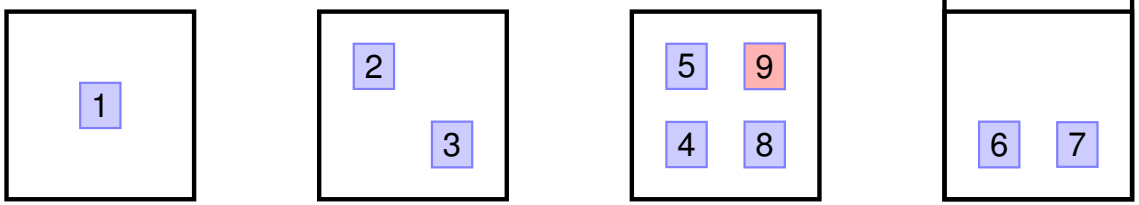


This beautiful plant is unique

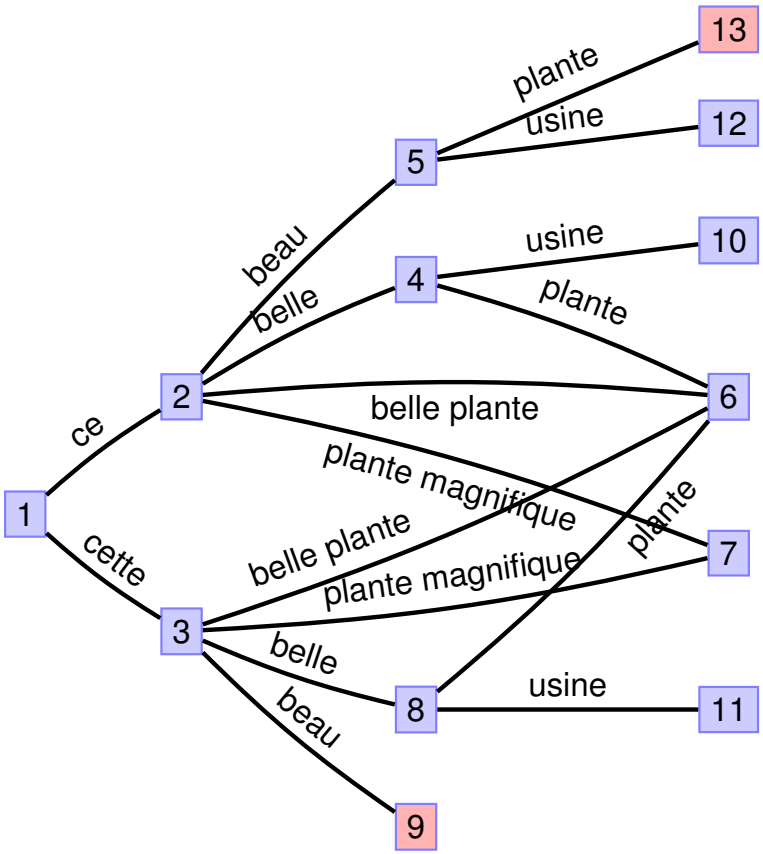


transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	

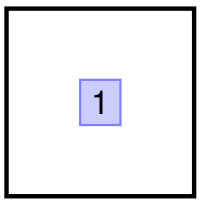


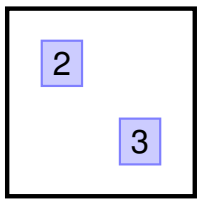
This beautiful plant is unique

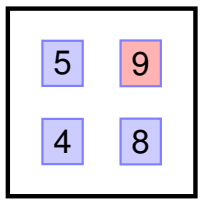


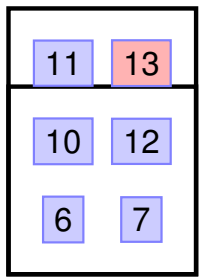
transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	

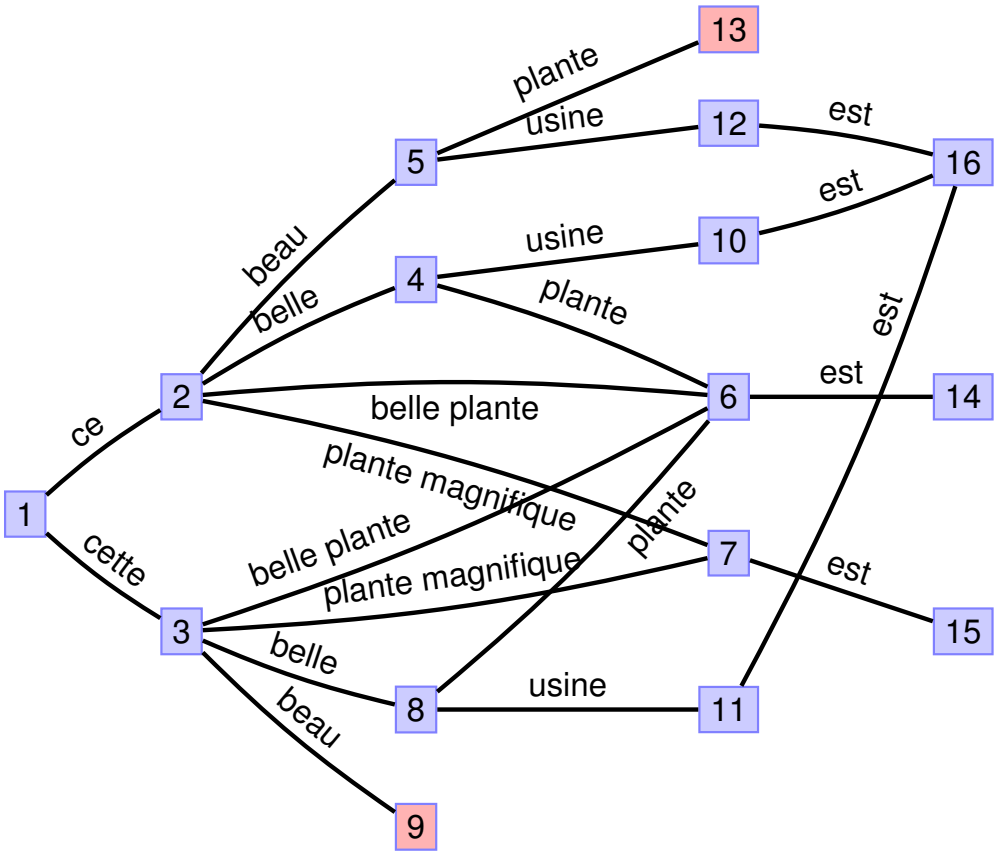






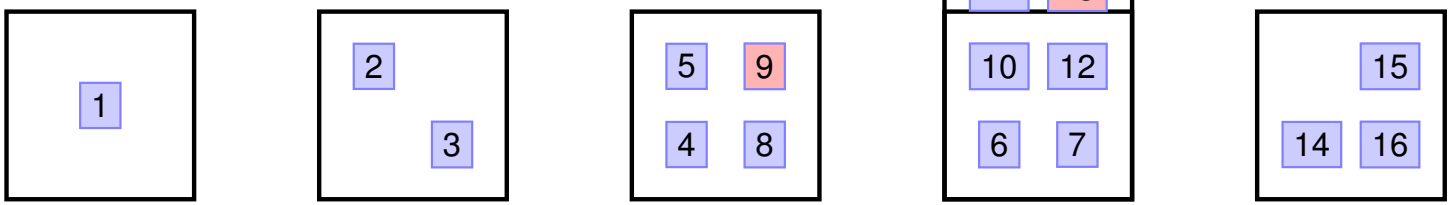


This beautiful plant is unique

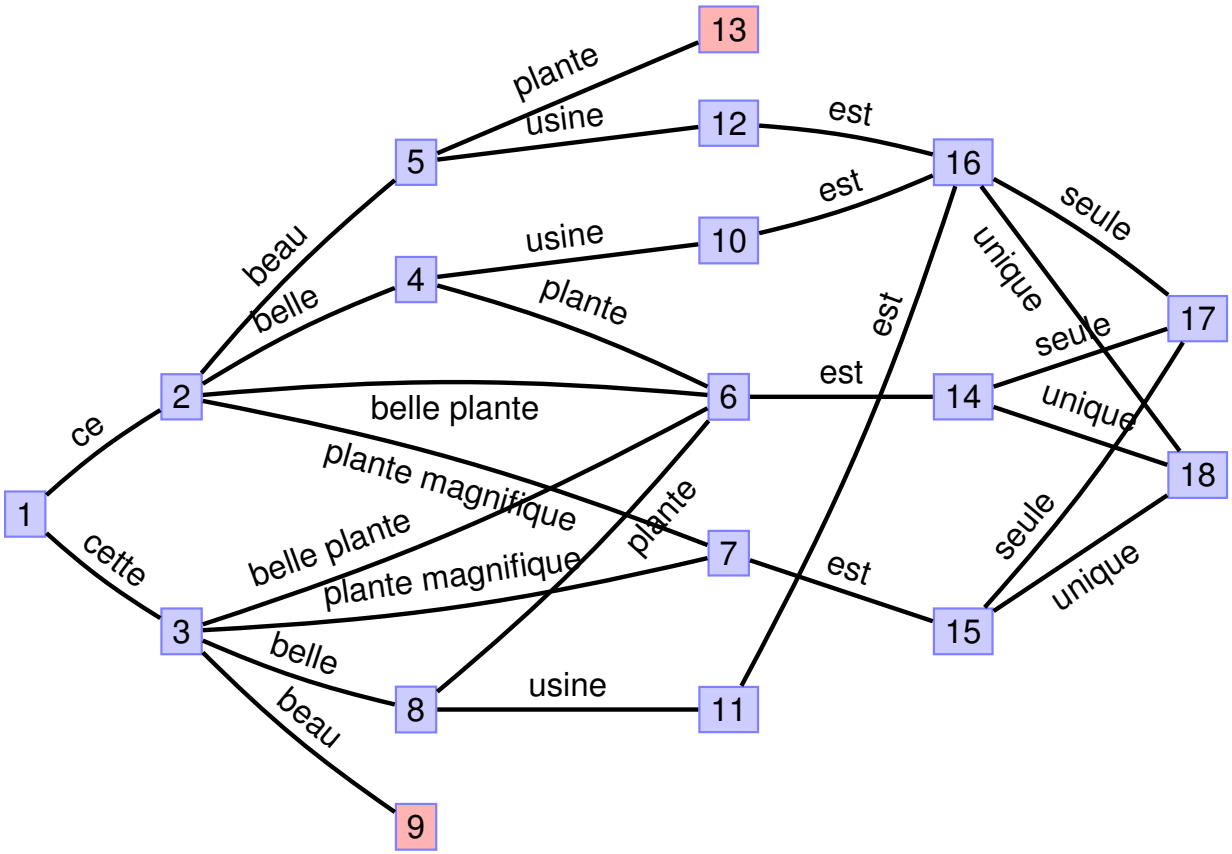


transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	

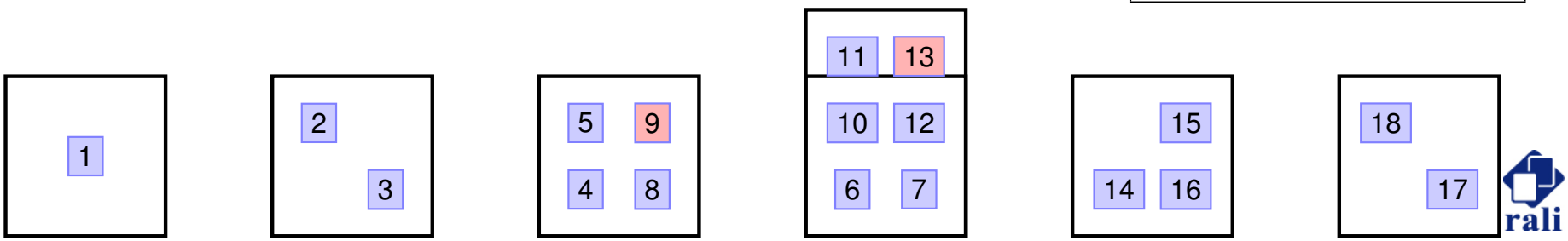


This beautiful plant is unique



transfer table		
this	↔	ce
	↔	cette
beautiful	↔	belle
	↔	beau
plant	↔	plante
	↔	usine
is	↔	est
unique	↔	seule
	↔	unique
beautiful plant ↕ belle plante plante magnifique		

language model	
ce beau plante	:-(
cette belle usine	:-|
belle usine est	:-)
. . .	



Multiple pass decoding

Motivations

- Some models are difficult to integrate into the first decoding stage
  - Higher order language models
  - Non cumulative scoring fonctions

$\Rightarrow$ Two steps approach :

1. Decoding a first list of likely candidates
2. Selecting the best candidate making use of extra information (selection, rescoring)

rali

An example of n-best list

```
Notre déclaration des droits est la première de ce millénaire .
 ||| lm: -53.17 tm: -8.55 -8.37 -6.30 -9.46 8.99907 w: -11 ||| -2.02
Notre déclaration des droits n ' est la première de ce millénaire .
 ||| lm: -55.96 tm: -4.29 -8.37 -5.71 -16.96 7.99 w: -13 ||| -2.11
Notre déclaration des droits est le premier de ce millénaire .#
 ||| lm: -52.68 tm: -8.69 -8.73 -7.27 -10.4078 8.99 w: -11 ||| -2.12
Notre déclaration des droits est la première de ce nouveau millénaire
 ||| lm: -53.42 tm: -10.69 -9.06 -9.48 -12.9981 8.99 w: -12 ||| -2.22
Notre déclaration des droits n ' est le premier de ce millénaire .
 ||| lm: -55.77 tm: -4.43 -8.73 -6.68 -17.9048 7.99 w: -13 ||| -2.23
Notre déclaration de droits est la première de ce millénaire .
 ||| lm: -59.42 tm: -3.33 -8.28 -5.19 -8.48052 7.99 w: -11 ||| -2.23
La déclaration des droits est la première de ce millénaire .
 ||| lm: -47.60 tm: -14.40 -14.26 -9.57 -12.6795 8.99 w: -11 ||| -2.3
Notre déclaration des droits n ' est la première de ce nouveau millén
 ||| lm: -56.20 tm: -6.433 -9.058 -8.89 -20.4951 7.99 w: -14 ||| -2.3
Notre déclaration des droits , c' est la première de ce millénaire .
 ||| lm: -54.70 tm: -9.60 -8.81 -10.84 -16.6753 8.99 w: -13 ||| -2.31
Notre déclaration des droits est la première de millénaire .
 ||| lm: -53.99 tm: -7.79 -12.01 -4.39 -8.24 6.99 w: -10 ||| -2.31
```

MERT $\equiv$ Minimum Error Rate Traning

[Och, 2003]

- Soit $D \equiv (\mathbf{f}_s, \mathbf{r}_s)_{s=1}^{S}$ un ensemble de phrases sources et leur traduction de référence et soit $\mathbf{C}_s \equiv \{\mathbf{e}_s^1, \ldots, \mathbf{e}_s^K\}$ un ensemble de K traductions de $\mathbf{f}_s$
- $E(D) = \sum_s E(\mathbf{r}_s, \mathbf{e}_s)$
- on cherche :

$$\begin{aligned} \hat{\lambda}_1^M &= \arg\min_{\lambda_1^M} \left\{ \sum_s E(\mathbf{r}_s, \hat{e}(\mathbf{f}_s, \lambda_1^M)) \right\} \\ &= \arg\min_{\lambda_1^M} \left\{ \sum_s \sum_k E(\mathbf{r}_s, \mathbf{e}_s^k) \delta(\hat{e}(\mathbf{f}_s, \lambda_1^M), \mathbf{e}_s^k) \right\} \end{aligned}$$

- où : $\hat{e}(\mathbf{f}_s, \lambda_1^M) = \operatorname{argmax}_{e \in \mathbf{C}_s} \{\sum_m \lambda_m h_m(\mathbf{e}|\mathbf{f}_s)\} \equiv \operatorname{argmax}_{e \in \mathbf{C}_s} \left\{ \lambda_1^M . \mathcal{H}_1^M(\mathbf{e}, \mathbf{f}_s) \right\}$ est le décodage (de $\mathbf{f}_s$) réalisé sous le régime λ_1^M

Note : la minimisation inclut un argmax ce qui interdit la descente de gradient $\Rightarrow$ *linesearch*

Line search dans [Och, 2003]

- Soit d une direction selon une des M dimensions de λ_1^M (ex : $d = (0, 0, 0, 1, 0, \ldots)$)
- $\lambda_1^M + \gamma d$ représente l'ensemble des déplacements dans la direction d si $\gamma \in \ ]-\infty, \infty[$
- chaque traduction $e \in \mathbf{C}_s$ est représentée par une droite qui représente son score en fonction de γ selon d :

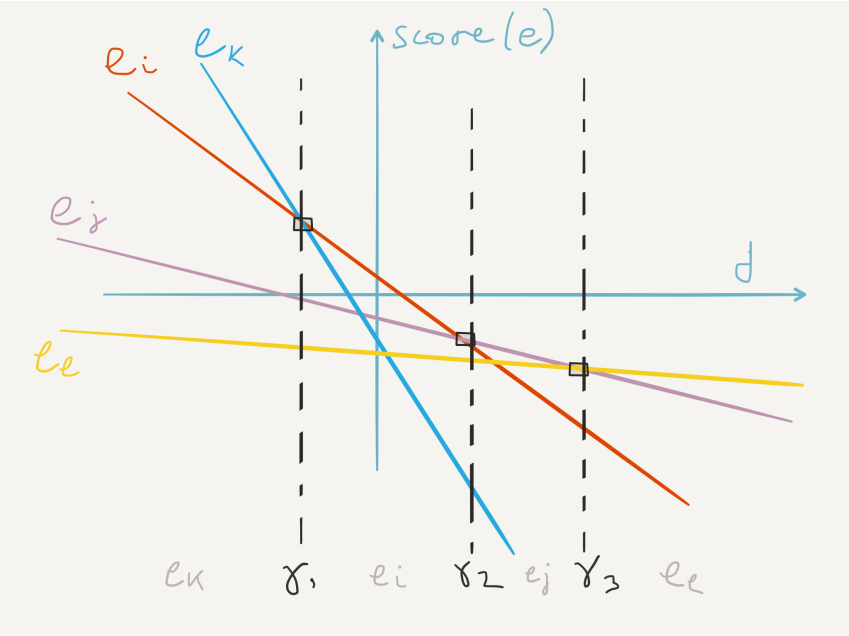
$$\begin{aligned} score(e, \gamma) &= (\lambda_1^M + \gamma d).\mathcal{H}_1^M(\mathbf{e}, \mathbf{f}_s) \\ &= \underbrace{\lambda_1^M.\mathcal{H}_1^M(\mathbf{e}, \mathbf{f}_s)}_{a(e)} + \gamma \ \underbrace{d\ .\mathcal{H}_1^M(\mathbf{e}, \mathbf{f}_s)}_{b(e)} \end{aligned}$$

  - note : $a(e)$ et $b(e)$ sont des constantes pour γ
- intuition : les intersections de ces droites indiquent des changements dans la fonction objectif (p.ex. BLEU). (la fonction objectif est linéaire par parties)

Line search dans [Och, 2003]

`linesearch(`$loss$, λ_1^M, d, $\{\mathcal{C}_1,\ldots,\mathcal{C}_S\}$`)` :

- $G \leftarrow \{\}$
- `foreach` s $\in [1, S]$
  - $G = G \cup$ inters. de $score(e_s^i, \gamma)$ avec $score(e_s^j, \gamma)$



- `return` $\arg\min_{\gamma \in G} loss(\{\mathcal{C}_1, \ldots, \mathcal{C}_S\}, \lambda_1^M + (\gamma - \epsilon).d)$

Line search dans [Och, 2003]

- $\lambda_1^M \leftarrow \{1, 1, \ldots\}$, i $\leftarrow$ 0
- *(1) :* tant que *pas marre* ou *plus rien à faire* :
  - décoder avec $\lambda_1^M \Rightarrow \mathcal{L}_i$ (en pratique $\mathcal{L}_i$ est ajouté à $\mathcal{L}_{i-1}$)
  - `foreach` d do `linesearch`(loss, λ_1^M, d, $\mathcal{L}_i$)
  - prendre d, γ qui minimise l'erreur (choix dans Moses) $\Rightarrow \lambda_1^M$
  - `++i`, aller en *(1)*

en pratique :

- sur un corpus de développement
- long (car décodage à chaque itération)
- bon pour des fonctions de scores de quelques features
- on parle de tuning du système

Tuning

- $\arg\min_{\lambda_1^M} \left\{ \sum_s \sum_k E(\mathbf{r}_s, \mathbf{e}_s^k) \frac{p(e_s^k)^\alpha}{\sum_k p(e_s^k)^\alpha} \right\}$ pour $\gamma \geq 1$ est une version lissée de cette fonction (on peut calculer le gradient) : $\arg\min_{\lambda_1^M} \left\{ \sum_s \sum_k E(\mathbf{r}_s, \mathbf{e}_s^k) \delta(\hat{e}(\mathbf{f}_s, \lambda_1^M), \mathbf{e}_s^k) \right\}$

- lire ***[Cer et al., 2008]*** pour des variantes de MERT et leur impact
- si trop de features (≥ 30), Batch MIRA ***[Cherry and Foster, 2012]*** est utilisé (disponible dans Moses)

rali

Modèles de traduction IBM

In : un *bitexte*, cad un corpus constitué de deux textes alignés au niveau des phrases. Le découpage en "mots" est également connu.

Out :

- modèle de transfert :

the	(3/149)	(le,0.18) (la,0.15) (de,0.12)
minister	(2/27)	(ministre,0.8) (le,0.12)
people	(3/66)	(gens,0.25) (les,0.16) (personnes,0.1)
years	(3/24)	(ans,0.38) (années,0.31) (depuis,0.12)

$$\forall s, \sum_t p(t|s) = 1$$

- autres distributions selon le modèle utilisé

How ? Entraînement par EM de modèles génératifs

$$P(F = f|E = e) = \sum_{a \in \mathcal{A}(e,f)} P(F = f, A = a|E = e)$$

$$a = (a_1, \ldots, a_m) \text{ avec } a_i \in [0, l] \; \forall i \in [1, m]$$

IBM 1 & 2

- jointe ($e \equiv e_1 \dots e_l$) :

$$P(f, a|e) = l(m|e) \prod_{j=1}^{m} a(a_j|j, m, e) t(f_j|e_{a_j})$$

avec $a(a_j|j, m, e) = 1/(l+1)$ dans le cas d'IBM1

- histoire générative :

1. Choisir m la longueur de la traduction f (selon l)
2. Générer indépendamment chaque mot f_j de f :
   1. choisir a_j une position ($\in [0, l]$) dans e associée à f_j (selon la distribution a). e_{a_j} est le mot qui est "responsable" de la génération de f_j
   2. choisir f_j sachant e_{a_j} (selon t)

Entraînement d'un modèle IBM1

[Brown et al., 1993]

- Soit un corpus $(f^s, e^s)\ \forall s \in [1, S]$,

$$c(\mathsf{f}|\mathsf{e}; e, f) = \frac{t(\mathsf{f}|\mathsf{e})}{\sum_{i=0}^{l} t(\mathsf{f}|e_i)} \overbrace{\sum_{j=1}^{m} \delta(f_j, \mathsf{f})}^{\text{nb de f dans } f} \overbrace{\sum_{i=0}^{l} \delta(\mathsf{e}, e_i)}^{\text{nb de e dans } e}$$

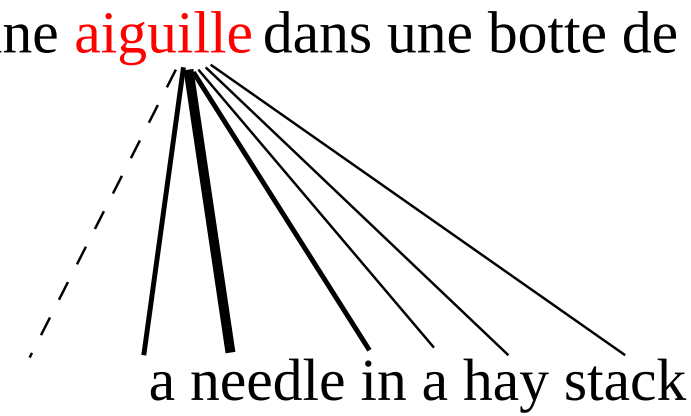
$$\text{avec } t(\mathsf{f}|\mathsf{e}) = \lambda_{\mathsf{e}} \sum_{s=1}^{S} c(\mathsf{f}|\mathsf{e}; f^{(s)}, e^{(s)})$$

- où $\delta(a, b)$ est la fonction de Kroneker qui vaut 1 lorsque $a = b$ et 0 sinon
- λ_e coefficient de Lagrange
- **Note :** converge vers un maximum global

rali

IBM2

- $a(i|j,l,m)$ probabilité qu'un mot source en position i soit associé à un mot cible en position j, sachant la longueur respective (comptée en mot) des deux phrases considérées.



- $\forall j,l,m \ \sum_{i=0}^{l} a(i|j,l,m) = 1$

Entraînement d'un modèle IBM2

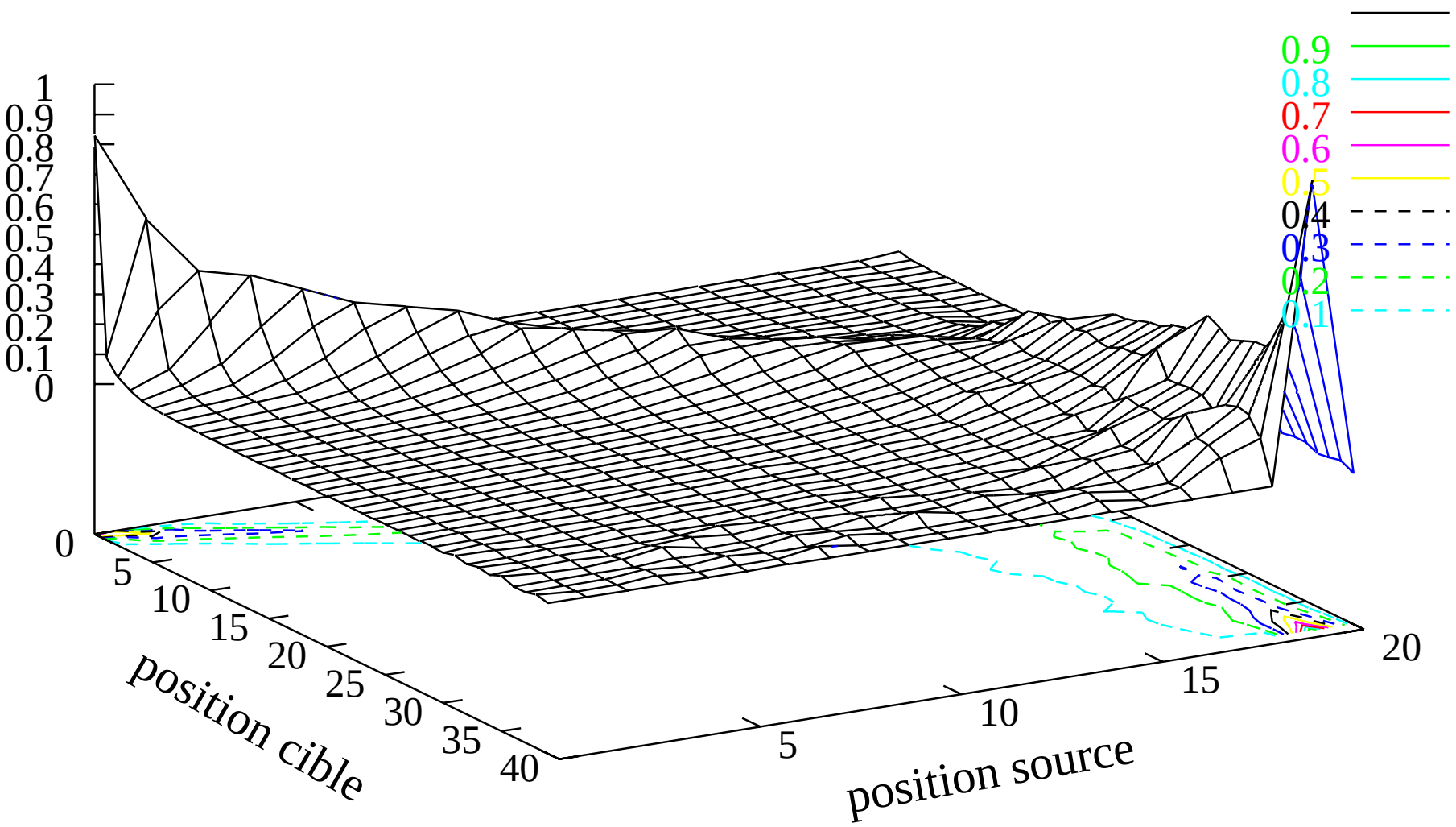
$$c(\mathsf{f}|\mathsf{e};e,f) = \sum_{j=1}^{m}\sum_{i=0}^{l}\frac{t(\mathsf{f}|\mathsf{e})a(i|j,l,m)}{\sum_{k=0}^{l}t(\mathsf{f}|e_k)a(k|j,l,m)}$$

$$c(i|j,l,m;f,e) = \frac{t(f_j|e_i)a(i|j,l,m)}{\sum_{k=0}^{l}t(f_j|e_k)a(k|j,l,m)}$$

$$t(\mathsf{f}|\mathsf{e}) = \lambda_{\mathsf{e}}\sum_{1}^{S}c(\mathsf{f}|\mathsf{e};e^s,f^s)$$

$$a(i|j,l,m) = \mu_{j,l,m}\sum_{1}^{S}c(i|j,l,m;f^s,e^s)$$

Probabilité d'alignement (IBM2)



Modèle HMM *[Vogel et al., 1996]*

- $p(a_j|a_{j-1}, l)$ permet de capturer le fait que les alignments sont (souvent) localement dépendants

Eve							x
avec						x	
pomme					x		
une				x			
mange		x	x				
Adam	x						
	Adam	is	eating	an	apple	with	Eve

- modélise la probabilité d'un saut : $p(a_j|a_{j-1}, l) = \frac{s(i-i')}{\sum_{k=1}^{l} s(k-i')}$

IBM3, 4 & 5 : intégration de la fertilité

- the est souvent traduit par un mot (le, la, l'), mais peut également ne pas être traduit.
- only est très souvent traduit par un seul mot (seulement), mais de temps en temps par deux mots (ne . . . que)
- bill est presque systématiquement traduit par 3 mots (projet de loi)

$\Rightarrow$ La fertilité modélise cette information : distribution du nombre de mots cibles générés par un mot source particulier :
$\forall \mathsf{e}, \sum_n \phi(n|\mathsf{e}) = 1$.

rali

IBM3, 4 & 5 : une autre histoire générative

L'histoire est simple, mais sa formulation mathématique et les problèmes calculatoires le sont moins ...

1. Choisir la fertilité de chaque mot source e_i selon la distribution $\phi(n|e_i) = \phi_i$.
2. Pour chaque mot source e_i, choisir ϕ_i mots cibles selon les distributions $t(\text{f}|\text{e})$. ***[Brown et al., 1993]*** appellent la liste des ϕ_i mots générés par e_i la *tablet* de e_i ; notée τ_i. L'ensemble des *tablets* est appelé le tableau de e.
3. Permuter les mots cibles selon une distribution Π afin d'obtenir la traduction.

IBM3, 4 & 5 : une autre histoire générative

- On peut voir cette histoire comme une suite de réécritures ***[Knight, 1999a]*** :

Mary did not slap the green witch	input
Mary not slap slap slap the the green witch	fertilité = $(1, 0, 1, 3, 2, 1$
Mary no daba una botefada a la verde bruja	cible
Mary no daba una botefada a la bruja verde	permutation
Mary no daba una botefada a la bruja verde	output

- l'histoire se complique un peu (au niveau des notations) avec ce que ***[Brown et al., 1993]*** appellent les spurious-words : les mots cibles non associés à un mot source particulier.

IBM3, 4 & 5 : une autre histoire générative

$$P(\tau, \pi|e) =$$

(step 1) $\prod_{i=1}^{l} P(\phi_i|\phi_1^{i-1}) \times P(\phi_0|\phi_1^l, e)$

(step 2) $\times \prod_{i=0}^{l} \prod_{k=1}^{\phi_i} P(\tau_{i,k}|\tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^l, e)$

(step 3) $\times \prod_{i=1}^{l} \prod_{k=1}^{\phi_i} P(\pi_{i,k}|\pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^{i-1}, \phi_0^l, e) \times \prod_{k=1}^{\phi_0} P(\pi_{0,k}|\pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, e)$

où $\tau_{i,1}^k = \tau_{i,1}, \tau_{i,2}, \ldots, \tau_{i,k}$ et $\pi_{i,1}^k = \pi_{i,1}, \ldots, \pi_{i,k}$.

$$P(f, a|e) = \sum_{(\tau,\pi)\in\langle f,a\rangle} P(\tau, \pi|e)$$

$\langle f, a\rangle$ désigne l'ensemble de toutes les paires (τ, π) consistantes avec la paire (f, a).

IBM3, 4 & 5 : une autre histoire générative

- τ et π nous donnent f et l'alignement sous-jacent.
- En général, différentes paires (τ, π) amènent à la même paire (f, a)

$$\tau_{cheap} = \{\text{bon}, \text{marché}\} \quad \pi_{cheap} = \{1, 2\}$$
$$\tau_{cheap} = \{\text{marché}, \text{bon}\} \quad \pi_{cheap} = \{2, 1\}$$

- Il y a $\phi_i!$ arrangements possibles des ϕ_i mots cibles de chaque ensemble τ_i, donc

$$|\langle f, a \rangle| = \prod_{i=0}^{l} \phi_i!$$

IBM3

$P(\phi_i|\phi_1^{i-1}, e)$ (pour $i \in [1, l]$) ne dépend que de $e_i \rightarrow n(\phi|\text{e})$, les probabilités de fertilité d'un mot source e.

$P(\tau_{i,k}|\tau_{i,1}^{k-1}, \tau_0^{i-1}, \phi_0^l, e)$ (pour $i \in [0, l]$) ne dépend que de $e_i \rightarrow t(\text{f}|\text{e})$, les probabilités de transfert.

$P(\pi_{i,k}|\pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^{i-1}, \phi_0^l, e)$ (pour $i \in [1, l]$) ne dépend que de i, m et $l \rightarrow d(j|i, m, l)$ les probabilités de distorsion.

- Les probabilités de distorsion et de fertilité pour e_0 sont traitées à part.
  - le nombre de mots cibles associés à e_0 (les *spurious words*) est choisi après avoir choisi les autres fertilités ($P(\phi_0|\phi_1^l, e)$)
  - le positionnement de ces mots est fait après avoir positionné les autres mots ($\prod_{k=1}^{\phi_0} P(\pi_{0,k}|\pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, e)$)

rali

IBM3

- les mots spurious sont placés de manière uniforme dans les espaces vacants, donc $p(\pi_{0,k} = j|\pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, e)$ vaut 0 si j n'est pas une place vacante et $(\phi_0 - k + 1)^{-1}$ sinon.
  - $\prod_{k=1}^{\phi_0} P(\pi_{0,k}|\pi_{0,1}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, e) = \prod_{k=1}^{\phi_0} \frac{1}{\phi_0 - k + 1} = \frac{1}{\phi_0!}$
- modélisé par une binomiale dépendant de p_1 :

$$P(\phi_0|\phi_1^l, e) = \binom{\phi = \phi_1 + \ldots + \phi_l}{\phi_0} (1 - p_1)^{\phi - \phi_0} (p_1)^{\phi_0}$$

où p_1 représente la probabilité que chaque mot des τ_1^l entraîne l'apparition d'un mot spurious

IBM3 : retour à l'histoire générative

1. pour chaque mot e de la phrase source, on choisi une fertilité ϕ avec une probabilité $n(\phi|\mathrm{e})$
2. m' est la somme de ces fertilités (n'incluant pas celle de e_0)
3. créer une nouvelle phrase source en retirant les mots de fertilité 0, en copiant les mots de fertilité 1, en dupliquant les mots de fertilité 2, etc.
4. après chacun de ces m' mots, décider si l'on insère un mot spurious (avec une probabilité p_1), ou pas.
5. Soit ϕ_0 le nombre de mots spurious finalement ajoutés. Alors $m = m' + \phi_0$ est la longueur de la traduction.
6. remplace chaque mot de cette phrase source par un mot cible selon les probabilités de transfert.
7. détermine la position de chaque mot cible non généré par e_0 selon les probabilités de distorsion.
8. si une position cible contient plus d'un mot, alors ERREUR
9. détermine les positions cibles des ϕ_0 mots générés par NULL parmi les positions restées vacantes dans f_1^m ; chaque choix étant équiprobable $(1/\phi_0)$.
10. lire la chaîne ainsi créée.

rali

IBM3 en un seul morceau

$$
\begin{aligned}
p(f|e) &= \sum_a p(f,a|e) \\
&= \sum_a \sum_{(\tau,\pi)\in\langle f,a\rangle} \\
&\quad \prod_{i=1}^{l} n(\phi_i|e_i) \times \binom{\phi=\phi_1+\ldots+\phi_l}{\phi_0}(1-p_1)^{\phi-\phi_0} p_1^{\phi_0} \times \\
&\quad \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} t(\tau_{i,k}|e_i) \times \\
&\quad \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} d(\pi_{i,k}|i,m,l) \times \frac{1}{\phi_0!} \\
&= \sum_a \binom{m-\phi_0}{\phi_0}(1-p_1)^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i!\, n(\phi_i|e_i) \times \\
&\quad \prod_{j=1}^{m} t(f_j|e_{a_j}) \prod_{j:a_j\neq 0}^{m} d(j|a_j,m,l)
\end{aligned}
$$

Note : lorsque (τ,π) est consistant avec (f,a) :

$$
\begin{aligned}
p(\tau|\phi,e) &= \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} t(\tau_{i,k}|e_i) &&= \prod_{j=1}^{m} t(f_j|e_{a_j}) \\
p(\pi|\tau,\phi,e) &= \frac{1}{\phi_0!}\prod_{i=1}^{l}\prod_{k=1}^{\phi_i} d(\pi_{i,k}|i,l,m) &&= \frac{1}{\phi_0!}\prod_{j:a_j\neq 0} d(j|a_j,l,m)
\end{aligned}
$$

rali

IBM3

Formules de réestimation (merci *[Brown et al., 1993]*) :

$$
\begin{aligned}
c(\mathsf{f}|\mathsf{e};f,e) &= \textstyle\sum_a p(a|e,f) \sum_{j=1}^{m} \delta(\mathsf{f},f_j)\delta(\mathsf{e},e_{a_j}) \\
c(j|i,m,l;f,e) &= \textstyle\sum_a p(a|e,f)\delta(i,a_j) \\
c(\phi|\mathsf{e};f,e) &= \textstyle\sum_a p(a|e,f) \sum_{i=1}^{l} \delta(\phi,\phi_i)\delta(e,e_i) \\
c(0;f,e) &= \textstyle\sum_a p(a|e,f)(m-2\phi_0) \\
c(1;f,e) &= \textstyle\sum_a p(a|e,f)\phi_0
\end{aligned}
$$

$$
\begin{aligned}
t(\mathsf{f}|\mathsf{e}) &\propto \textstyle\sum_{s=1}^{S} c(\mathsf{f}|\mathsf{e};f^s,e^s) \\
d(j|i,m,l) &\propto \textstyle\sum_{s=1}^{S} c(j|i,m,l;f^s,e^s) \\
n(\phi|\mathsf{e}) &\propto \textstyle\sum_{s=1}^{S} c(\phi|\mathsf{e};f^s,e^s) \\
p_k &\propto \textstyle\sum_{s=1}^{S} c(k;f^s,e^s)
\end{aligned}
$$

rali

IBM3

- **Problème :** On ne peut pas calculer efficacement la sommation sur l'ensemble des alignments ni l'alignement optimal selon le modèle IBM3
- *[Brown et al., 1993]* calculent cette somme sur un ensemble d'alignement "prometteurs" :
  - alignement le plus probable selon un modèle calculable (IBM2)
  - amélioration selon IBM3 sur les alignements obtenus par un ensemble restreint de modifications de ce "meilleur" alignement (approche greedy)
  - deux opérations élémentaires : *move* et *swap*.

rali

IBM3 : La magie des nombres (pris de *[Brown et al., 1993]*)

Probabilités de transfert et de fertilité pour le mot e =nodding.

f	$t(\text{f}\|\text{e})$	f	$t(\text{f}\|\text{e})$	ϕ	$n(\phi\|\text{e})$
signe	0.164	hocher	0.048	4	0.342
la	0.123	faire	0.030	3	0.293
tête	0.097	me	0.024	2	0.167
oui	0.086	approuve	0.019	1	0.163
fait	0.073	qui	0.019	0	0.023
que	0.073	un	0.012		
hoche	0.054	faites	0.011		

IBM3 : La magie des nombres (pris de *[Brown et al., 1993]*)

Probabilités de transfert et de fertilité pour le mot e =farmers.

f	$t(\text{f}\|\text{e})$	ϕ	$n(\phi\|\text{e})$
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		

Une propriété de IBM3

- Le fait que les probabilités de distorsion (la probabilité d'une position cible) ne dépendent pas des positions décidées auparavant fait que le modèle est déficient.
- En clair, cela signifie qu'il se peut très bien que plusieurs mots soient attribués indépendamment à la même position cible (ce que les modèles IBM interdisent par essence). Donc une partie de la masse de probabilité du modèle 3 est employée à modéliser des séquences sans intérêt.

$$\sum_f p(f|e) + p(failure|e) = 1$$

- Un modèle est déficient si $p(failure|e) > 0$

IBM4

- tendance à préserver localement l'ordre des mots

$\Rightarrow$ changement de l'histoire générative

- chaque mot source e_i génère toujours ϕ_i mots cibles (fertilité), mais on commence (**étape 1**) par générer la tête de la *tablet* (le premier mot) en tenant compte de la *tablet* précédente selon la distribution :

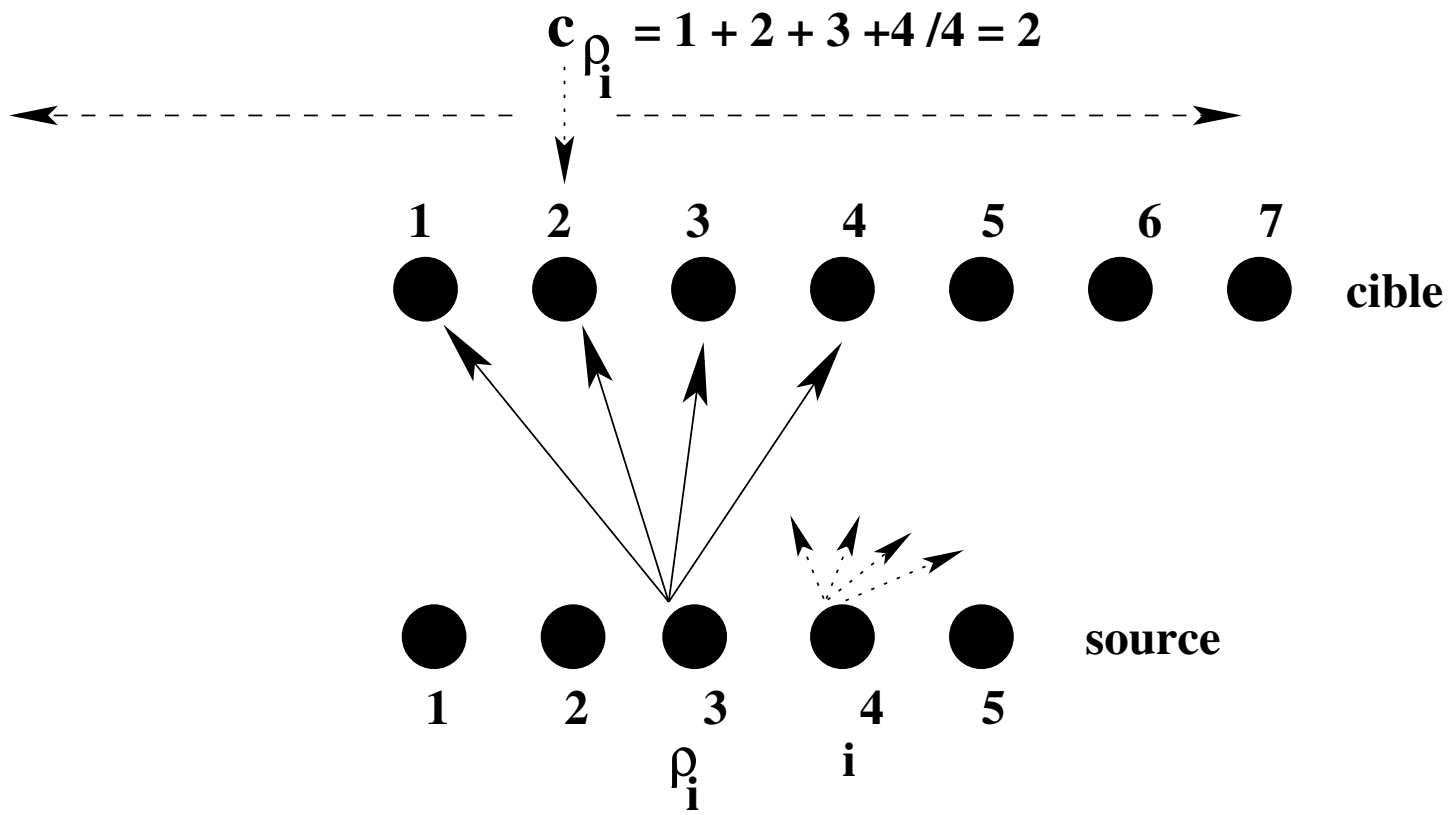
$$(\pi_{i,1} = j|\pi_1^{i-1}, \tau_0^{i-1}, \phi_0^l, e) = d_1(j - c_{\rho_i}|\mathcal{A}(e_{\rho_i}), \mathcal{B}(f_j))$$

avec :

$$\begin{aligned} \rho_i &= \max_{i'<i}\{i' : \phi_{i'} > 0\} \\ c_\rho &= \phi_\rho^{-1} \sum_{k=1}^{\phi_\rho} \pi_{\rho,k} \end{aligned}$$

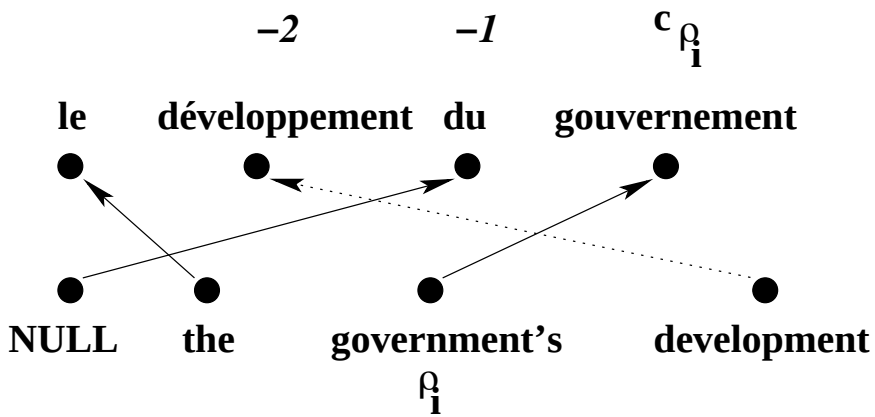
et $\mathcal{A}$ et $\mathcal{B}$ deux fonctions prenant respectivement en entrée le mot source "précédent" et le mot de tête (cible) de la i-ème tablet $\tau_{i,1}$.

IBM4 : Illustration de l'étape 1



IBM4

- L'idée est de modéliser les mouvements de groupes par ces déplacements de têtes (qui peuvent être positifs ou négatifs).
- permet de capturer des régularités positionnelles entre les deux langues comme par exemple la tendance à l'inversion (entre le français et l'anglais) de la séquence nom/adjectif (ex : The blue montain / la montagne bleue).
- ***[Brown et al., 1993]*** rapportent que :
  $d_1(-1|\mathcal{A}(\text{government's})\mathcal{B}(\text{développement})) = 0.7986$ et
  $d_1(+1|\mathcal{A}(\text{government's})\mathcal{B}(\text{développement})) = 0.0168$.



IBM4

- Une fois la tête du *tablet* positionnée, on positionne les autres mots (**étape 2**) en imposant que chaque mot positionné se trouve à droite du mot précédemment positionné :

$$p(\pi_{i,k} = j|\pi_{i,1}^{k-1}, \pi_1^{i-1}, \tau_0^{i-1}, \phi_0^l, e) = d_{>1}(j - \pi_{i,k-1}|\mathcal{B}(f_j))$$

- ***[Brown et al., 1993]*** rapportent par exemple que $d_{>1}(2|\mathcal{B}(\text{pas})) = 0.6847$ (ex : not/ne... pas) alors que $d_{>1}(2|\mathcal{B}(\text{en})) = 0.15$ (ex : implemented/mis en application)
- IBM4 est déficient.

IBM5

- IBM3 est déficient en raison de l'indépendance du choix d'une position cible pour un mot donné d'une *tablet*.
- IBM4 possède (d'une manière moindre) les mêmes travers, plus celui de modéliser des déplacements de tête de tablet en dehors de la phrase source (avant le début ou après la fin).
- IBM5 remédie à cela en forçant les positionnements à tomber sur des places vacantes. Pour plus de détails, voir ***[Brown et al., 1993]***. Le modèle 5 contient plus de paramètres, mais ne donne pas nécessairement de meilleurs résultats que IBM4.

Inversion Transduction Grammar (ITG)

- ***[Wu, 1997, Wu, 2000]***
- grammaire hors contexte qui génère un bitexte
- règles de la forme :

  $A \rightarrow (x/y \quad \vee \quad B)^+$

  où x/y est une paire de mots (ex : apple/pomme), A et B des non-terminaux.
- Les paires lexicales peuvent contenir le mot vide ϵ pour rendre compte des insertions de mots sources et cibles.

  Ex : $A \rightarrow B \; x/y \; C \; z/\epsilon$ signifie que x (source) se réécrit par y (cible) et z se réécrit en ϵ (insertion d'un mot source).

rali

ITG

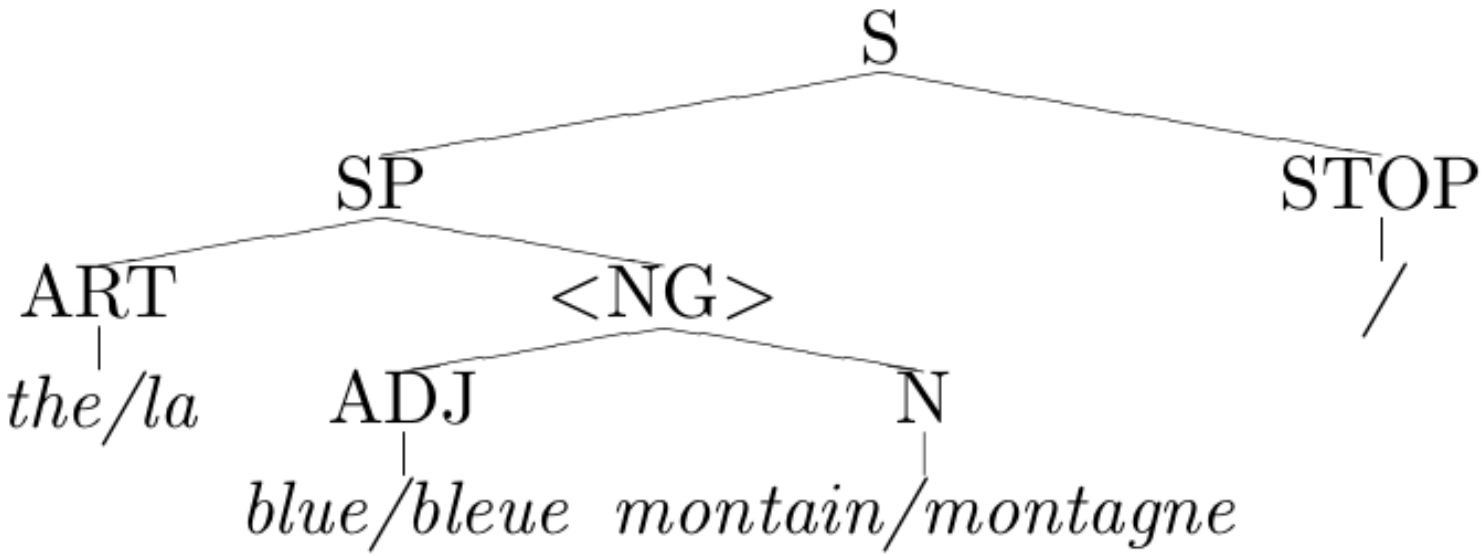
$S \rightarrow SP\ Stop$	$STOP \rightarrow ./.$
$SP \rightarrow ART\ NG$	$ART \rightarrow the/la$
$NG \rightarrow N\ ADJ$	$ADJ \rightarrow blue/bleue$
$NG \rightarrow ADJ\ N$	$N \rightarrow montain/montagne$

- Cette grammaire ne peut pas produire un alignement (correct) de la paire de phrases (the blue montain . / la montagne bleue .).
- ***[Wu, 2000]*** introduit deux types de lecture d'une règle :

  [] indique une lecture habituelle (gauche-droite)
  <> indique un lecture inversée.

$$NG \rightarrow [ADJ\ N]$$
$$NG \rightarrow \langle ADJ\ N \rangle$$

ITG



- Les fils directs des nœuds étiquetés <> sont lus de la gauche vers la droite dans la langue source et de la droite vers la gauche dans la langue cible.

ITG

- ***[Wu, 2000]*** montre qu'une ITG peut-être représentée sous forme normale :

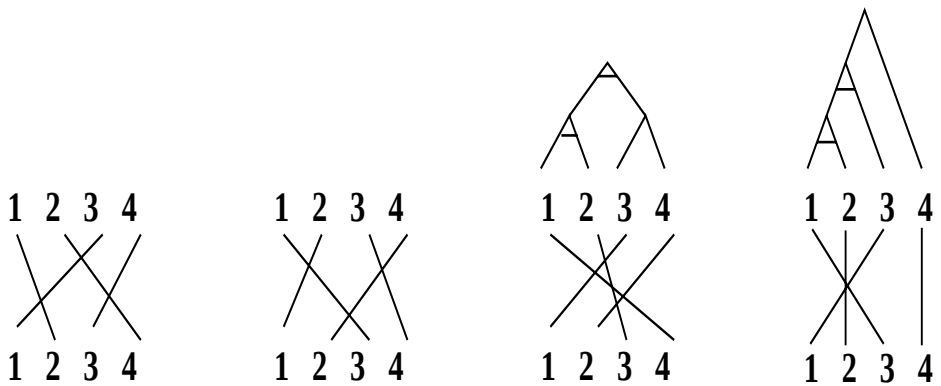
$$\begin{array}{lll} S \rightarrow \epsilon / \epsilon & A \rightarrow x/\epsilon & A \rightarrow [B\ C] \\ A \rightarrow x/y & A \rightarrow \epsilon /y & A \rightarrow \langle B\ C \rangle \end{array}$$

- Tous les alignements ne sont pas représentables avec une ITG. ***[Wu, 2000]*** fait l'hypothèse que ces alignements non représentables ne sont pas linguistiquement motivés.
- Si tel est le cas, alors une ITG possède l'intéressante propriété d'autoriser une fraction de plus en plus restreinte des alignements potentiels entre des constituants de longueur l, plus l augmente.

ITG

l	ITG	|aligt|	l	ITG	|aligt|
0	1	1	6	394	720
1	1	1	7	1806	5040
2	2	2	8	8558	40320
3	6	6	9	41586	362880
4	22	24	10	206098	3628800
5	90	120	...	...	...

- 2 alignements de séquences de 4 mots non représentables par une ITG (parmi 24 alignements potentiels) et 2 alignements représentables :



ITG

- version stochastique des ITG en ajoutant une probabilité à chaque règle sous la contrainte, pour tout non terminal A :

$$\sum_{B,C \in N} (a_{\text{A}\rightarrow\text{[B C]}} + a_{\text{A}\rightarrow\text{<B C>}}) + \sum_{x,y \in T} b_A(x,y) = 1$$

où les a-probabilités sont celles associées aux nœuds qui ne sont pas des feuilles, et les b-probabilités sont les probabilités lexicales.

- un algorithme d'analyse proche de CYK est décrit, ainsi que différents usages des ITGs comme le parenthésage bilingue à l'aide d'une ITG simple comme :

A → [A A] [a]	A → u_i/v_j [$b_{i,j}$]
A → <A A> [a]	A → u_i/ϵ [$b_{i,\epsilon}$]
	A → ϵ/v_j [$b_{\epsilon,j}$]

Traduction guidée par la syntaxe

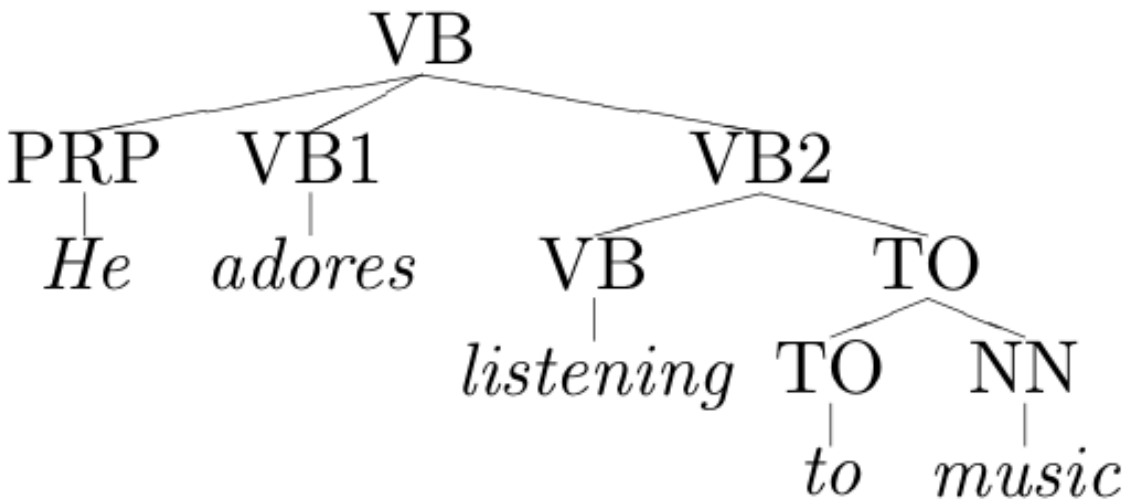
Canal bruité dans *[Yamada and Knight, 2001]* :

- Un arbre syntaxique entre dans un canal qui le transmet sous la forme d'une chaîne cible traduction de la chaîne source associée à l'arbre "entrant".
- Le canal applique pour cela trois opérations :
  1. réordonnancement des nœuds de l'arbre,
  2. insertion de mots dans les nœuds de l'arbre réordonné et,
  3. traduction des feuilles de l'arbre précédent.
- traduction = la chaîne résultant de la lecture en profondeur d'abord de l'arbre "traduit".

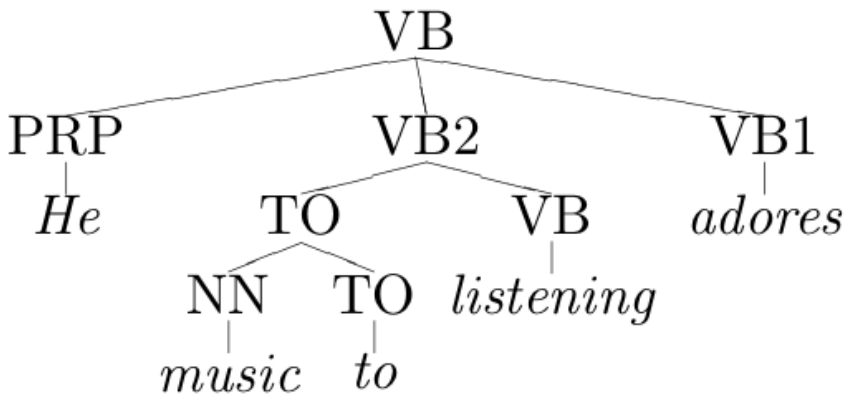
rali

[Yamada and Knight, 2001] : entrée

- arbre obtenu par l'analyseur probabiliste de *[Collins, 1999]*

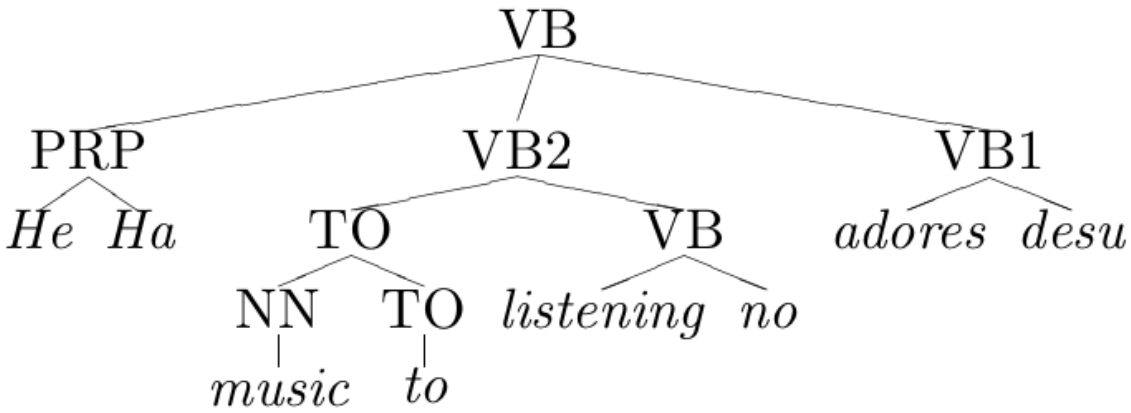


[Yamada and Knight, 2001] : Réordonnancement



original	réordonnanct	prob	original	réord.	prob
	PRP VB2 VB1	0.723	VB TO	TO VB	0.749
PRP VB1 VB2	VB2 PRP VB1	0.083		VB TO	0.251
	PRP VB1 VB2	0.074	TO NN	NN TO	0.893
	⋮	⋮		TO NN	0.107

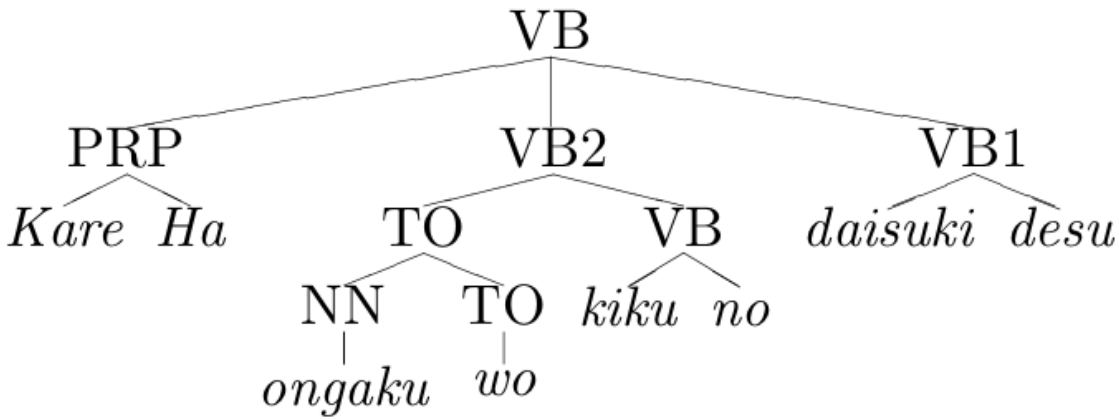
[Yamada and Knight, 2001] : Insertion



parent node	TOP VB	VB VB	VB PRP	... ...
p(none)	0.735	0.687	0.344	...
p(left)	0.004	0.061	0.004	...
p(right)	0.260	0.252	0.652	...

w	$p_{ins}(w)$
ha	0.219
ta	0.131
⋮	⋮

[Yamada and Knight, 2001] : Transfert



E	adores		he		...
J	daisuki	1.0	kare	0.952	...
			NULL	0.016	...
			nani	0.005	...
			da	0.003	...
			:	:	...

Les systèmes hiérarchiques *[Chiang, 2007]*

Jean donne une balle à Marie	|||	John gives Mary a ball
une balle	|||	a ball
Jean	|||	John
Marie	|||	Mary

Les systèmes hiérarchiques *[Chiang, 2007]*

Jean donne une balle à Marie	|||	John gives Mary a ball
une balle	|||	a ball
Jean	|||	John
Marie	|||	Mary

Jean donne X_1 à Marie	|||	John gives Mary X_1
X_1 donne une balle à Marie	|||	X_1 gives Mary a ball
Jean donne une balle à X_1	|||	John gives X_1 a ball
X_1 donne une balle à X_2	|||	X_1 gives X_2 a ball
X_1 donne X_2 à X_3	|||	X_1 gives X_2 X_3

Les systèmes hiérarchiques

$$G : \{N \equiv \{S, X\}, V, S, R, P\}$$

$$R : \left\{ \begin{array}{l} \{X \rightarrow \langle \delta, \gamma, \sim \rangle \ : \ \delta, \gamma \in (N \cup V)^{\star}\} \\ S \rightarrow \langle SX, SX, 1-2 \rangle \\ S \rightarrow \langle X, X, 1 \rangle \end{array} \right\}$$

$$P : p(r \equiv X \rightarrow \langle \delta, \gamma, \sim \rangle) = \prod_i \phi_i(r)^{\lambda_i}$$

Décodage : $trad(f) \approx \text{argmax}_{D:yield(D)\equiv f}\, w(D)$ où : $w(D) = \prod_{r \in D} p(r)$

Traduction = analyse

Les systèmes hiérarchiques

$S \rightarrow \langle SX, SX \rangle$	$S \rightarrow \langle X, X \rangle$
$X \rightarrow \langle \text{John}, \text{Jean} \rangle$	$X \rightarrow \langle \text{saw } X_1 \text{ with } X_2, \text{à vu } X_1 \text{ avec } X_2 \rangle$
$X \rightarrow \langle \text{the man}, \text{l'homme} \rangle$	$X \rightarrow \langle \text{saw } X_1, \text{a vu } X_1 \rangle$
$X \rightarrow \langle \text{with a } X_1, \text{avec un } X_1 \rangle$	$X \rightarrow \langle \text{telescope}, \text{télescope} \rangle$

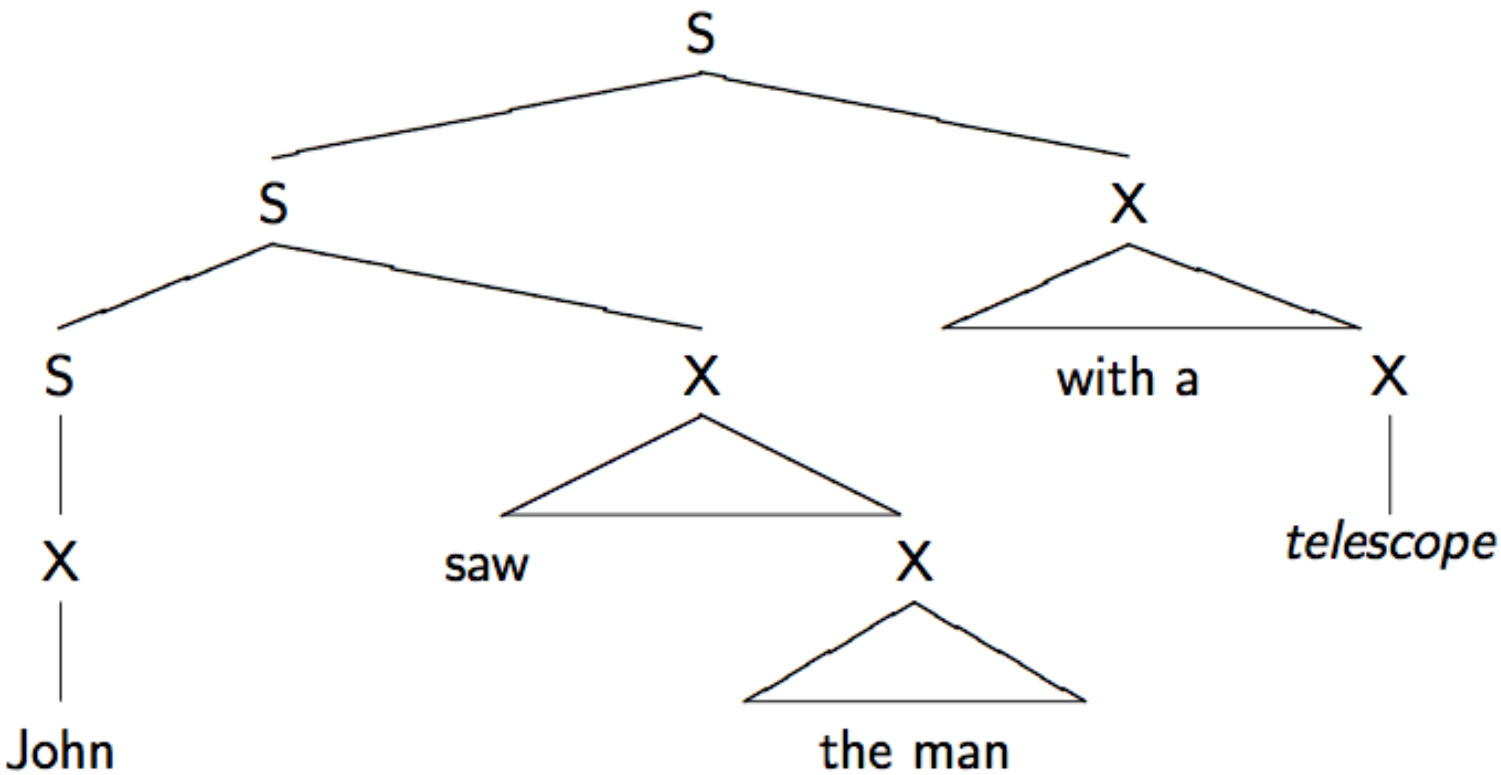
Les systèmes hiérarchiques

$S \rightarrow \langle SX, SX \rangle$	$S \rightarrow \langle X, X \rangle$
$X \rightarrow \langle \text{John}, \text{Jean} \rangle$	$X \rightarrow \langle \text{saw } X_1 \text{ with } X_2, \text{à vu } X_1 \text{ avec } X_2 \rangle$
$X \rightarrow \langle \text{the man}, \text{l'homme} \rangle$	$X \rightarrow \langle \text{saw } X_1, \text{a vu } X_1 \rangle$
$X \rightarrow \langle \text{with a } X_1, \text{avec un } X_1 \rangle$	$X \rightarrow \langle \text{telescope}, \text{télescope} \rangle$

$$
\begin{aligned}
S &\Rightarrow \langle SX, SX \rangle \\
&\Rightarrow \langle SX_1X_2, SX_1X_2 \rangle \\
&\Rightarrow \langle X_1X_2X_3, X_1X_2X_3 \rangle \\
&\Rightarrow \langle \text{John } X_1X_2, \text{Jean } X_1X_2 \rangle \\
&\Rightarrow \langle \text{John saw } X_1X_2, \text{Jean a vu } X_1\ X_2 \rangle \\
&\Rightarrow \langle \text{John saw the man } X_1, \text{Jean a vu l'homme } X_1 \rangle \\
&\Rightarrow \langle \text{John saw the man with a } X_1, \text{Jean a vu l'homme avec un } X_1 \rangle \\
&\Rightarrow \langle \text{John saw the man with a telescope} , \\
&\qquad \text{Jean a vu l'homme avec un télescope} \rangle
\end{aligned}
$$

Les systèmes hiérarchiques

$S \rightarrow \langle SX, SX \rangle$	$S \rightarrow \langle X, X \rangle$
$X \rightarrow \langle John, Jean \rangle$	$X \rightarrow \langle saw\ X_1\ with\ X_2, à\ vu\ X_1\ avec\ X_2 \rangle$
$X \rightarrow \langle the\ man, l'homme \rangle$	$X \rightarrow \langle saw\ X_1, a\ vu\ X_1 \rangle$
$X \rightarrow \langle with\ a\ X_1, avec\ un\ X_1 \rangle$	$X \rightarrow \langle telescope, télescope \rangle$



Désambiguïsation lexicale

- (Schwenk,2007 ; Stroppa et al, 2007 ; Carpuat et Wu, 2007)

En : You must make the first move.

En : You must first move the car.

Désambiguïsation lexicale

- (Schwenk,2007 ; Stroppa et al, 2007 ; Carpuat et Wu, 2007)

En : You must make the first move.

Fr : Tu dois faire le premier pas.

En : You must first move the car.

Fr : Tu dois d'abord déplacer la voiture.

Désambiguïsation lexicale

- (Schwenk,2007 ; Stroppa et al, 2007 ; Carpuat et Wu, 2007)

En : You must make the first move.

Fr : Tu dois faire le premier pas.

En : You must first move the car.

Fr : Tu dois d'abord déplacer la voiture.

Traduction de *first move* ? ⟨ premier pas / d'abord déplacer

Désambiguïsation lexicale

- (Schwenk,2007 ; Stroppa et al, 2007 ; Carpuat et Wu, 2007)

En : You must make the first move.
PP MD VV DT JJ NN

Fr : Tu dois faire le premier pas.

En : You must first move the car.
PP MD RB VV DT NN

Fr : Tu dois d'abord déplacer la voiture.

Les catégories lexicales permettent de désambiguïser

Désambiguïsation lexicale

- (Schwenk,2007 ; Stroppa et al, 2007 ; Carpuat et Wu, 2007)

En : You must make the first move.
PP MD VV DT JJ NN

Fr : Tu dois faire le premier pas.

En : You must first move the car.
PP MD RB VV DT NN

Fr : Tu dois d'abord déplacer la voiture.

Traductions : $\text{move}_{NN} \rightarrow$ pas
$\text{move}_{VV} \rightarrow$ déplacer

Utilisation de morpho-syntaxe

Principe

- Étiqueter les textes parallèles avec des informations morpho-syntaxiques
- Enrichir les mots avec les catégories lexicales :

$$You_P\ must_V\ make_V\ the_D\ first_{Adj}\ move_N.$$
$$Tu_P\ dois_V\ faire_V\ le_D\ premier_{Adj}\ pas_N.$$

- Construire un **système statistique** complet sur ce vocabulaire enrichi
- En sortie :
  - Suppression des étiquettes
  - Réutilisation des étiquettes (ML morpho-syntaxique)

rali

Modèle de Traduction Factorisé
[Koehn and Hoang, 2007]

Motivation

- Seuls sont disponibles les segments du corpus parallèle d'apprentissage
- Pas de généralisation lexicale

Modèle de Traduction Factorisé
[Koehn and Hoang, 2007]

Motivation

- Seuls sont disponibles les segments du corpus parallèle d'apprentissage
- Pas de généralisation lexicale

Exemple

- *La voiture rouge est belle*
  $\rightarrow$ *The red car is nice*
- *Les vélos rouges sont beaux*
  $\rightarrow$ *The red bikes are nice*
- *Les voitures rouges sont belles*
  Traduction de cette phrase sachant les deux autres ?

Modèle de Traduction Factorisé

Principe

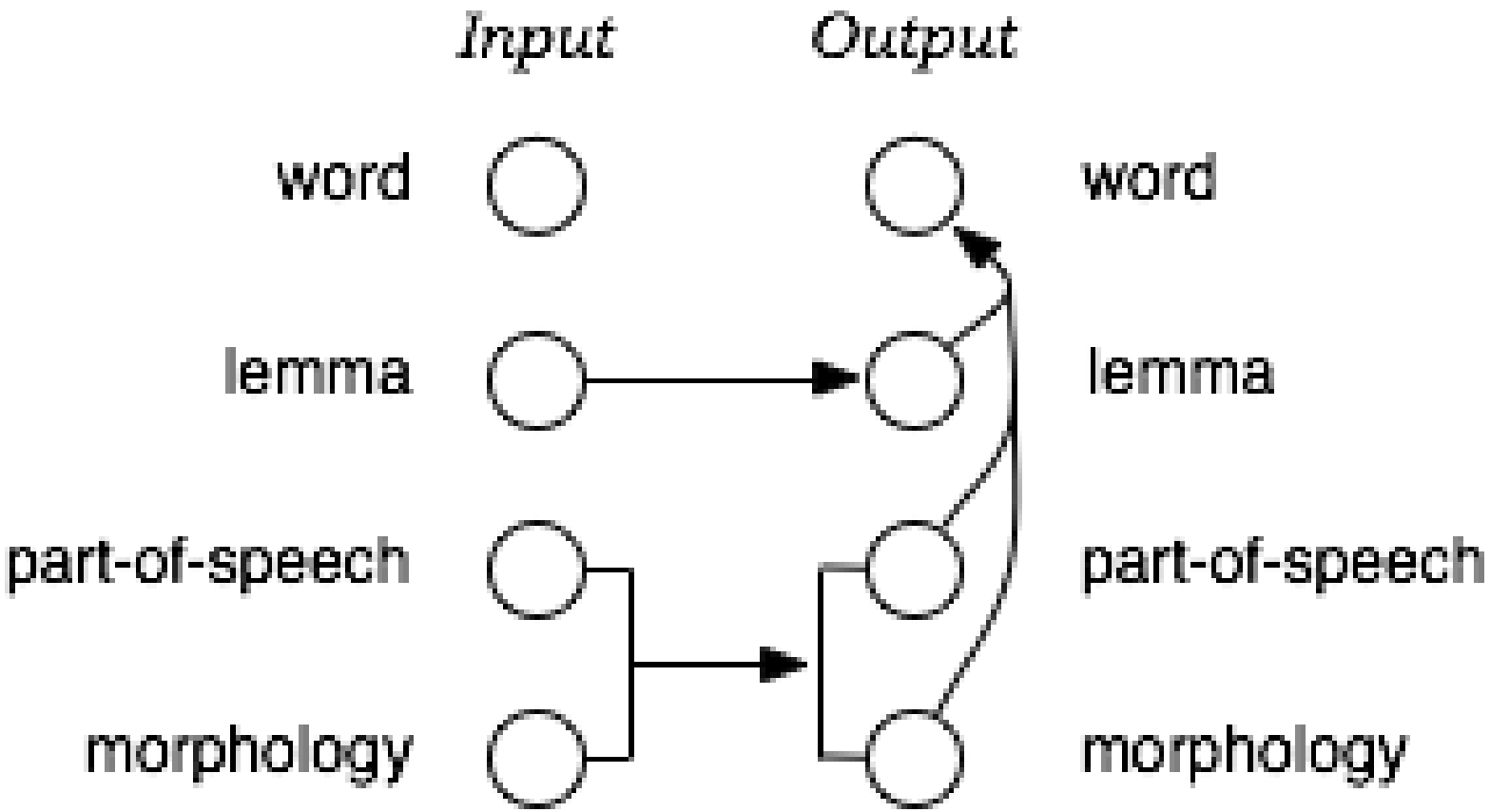
- L'approche actuelle de traduction par syntagmes traite un mot comme unité
- Savoir traduire un mot, ne permet pas de traduire son pluriel, sa conjugaison, ...

$\Rightarrow$ Décomposer les mots en lemme, genre, nombre, ...

- Traduire ces facteurs séparément
- Recomposer le mot dans la langue cible à partir de la traductions des facteurs
- Processus de génération

rali

Modèle de Traduction Factorisé : Mise en œuvre



Le système Systran

In **1962**, there were 48 working groups deeply involved in the research and development of MT.
The US National Academy of Sciences published the **ALPAC report** in **1966**. This report stated that MT had a limited future, and then put an end to all US Government research and development projects and financing for MT.
Peter Toma, Ph.D., a linguist researcher for MT, began his work in **1957** at the California Institute of Technology. Later, Dr. Toma became involved in the pioneering work in Russian to English MT at Georgetown University, the largest MT project in the US of that time. In **1968**, Dr. Toma established a company in La Jolla, California, USA, with a product called SYSTRAN ; an acronym for System Translation. Soon thereafter, the company was contracted to develop Russian to English MT for the US Air Force.
The first SYSTRAN system was tested in early **1969** at Wright-Patterson Air Force Base in Dayton, Ohio, USA. Since **1970**, the system has continued to provide translations for the US Air Force's Foreign Technology Division.
During the period **1974-1975**, SYSTRAN was used by NASA for the joint US-USSR Apollo-Soyouz space project.

texte consulté en 2002 depuis : `http://www.translation.net/systrans.html`

rali

Systran

In **1975**, Dr. Toma demonstrated a prototype of English to French MT to representatives of the Commission of the European Communities (CEC), which resulted in a contract to develop MT systems for various European language pairs. The CEC uses more than 12 SYSTRAN MT systems for the translation of its internal documents.
Xerox Corporation began using SYSTRAN in **1978**, the same period during which the SYSTRAN MT systems were transformed into multitarget (English to several non-English languages). Xerox continues to use the systems for translation of thousands of pages per year. This process allows Xerox to launch multilingual products to the global marketplace.
In **1981**, SYSTRAN initiated the development of the Japanese to English and English to Japanese MT systems. Under newly European influence, the first World SYSTRAN Conference, organized by the CEC, was held in Luxembourg, shortly thereafter. This conference, the first dedicated to one single MT system, brought together all the principal SYSTRAN users from around the world.
SYSTRAN introduced the utility "Customer Specific Dictionaries", (referred to as CSDs), in **1989**. CSDs are dictionaries created by users with their own specific terminology.
The company developed integrated MT for Xerox in **1990**. This new system provided the option to preserve the original document format of a text during the translation process. This option has

Systran

In **1992**, SYSTRAN began the ”C” conversion project, bringing its powerful, patented MT technology to the PC.
In **1995**, SYSTRAN PROfessional for Windows in standalone and client/server versions were launched.
SYSTRAN received a contract from the US National Air Intelligence Center to develop several Eastern European MT language pairs in **1996**. This included the first-ever Serbo-Croatian to English MT system, already delivered to the US Government. Also in 1996, SYSTRAN signed a licensing agreement with Seiko Instruments, Inc., through which SYSTRAN would provide linguistic data and software for Seiko’s hand-held translation products.
In early **1997**, Ford Motor Company acquired multiple custom software licenses of SYSTRAN’s MT software, to be embedded directly into Ford’s systems.
Today, **14 SYSTRAN MT language pairs** are commercially available*

*52 paires de langues en février 2015

rali

Systran de l'intérieur (*[Hutchins and Somers, 1992]*)

- Consultation de dictionnaires bilingues à grande couverture.
  - dictionnaire principal dont les entrées sont des mots seuls
  - différents dictionnaires spécialisés (ou contextuels) dont les entrées sont des unités spécialisées.
- Une entrée dans le dictionnaire principal contient de nombreuses informations morphologiques, grammaticales et sémantiques.
  - valence d'un verbe,
  - type de nom (animé, énumérable, abstrait, etc.), etc.
- Le dictionnaire principal est lemmatisé (sauf pour l'anglais).
  - à chaque entrée lemmatisée correspond une association privilégiée dans la langue cible, elle même étiquetée avec des informations nécessaires à l'étape de génération.

rali

Systran de l'intérieur

Les dictionnaires contextuels contiennent entre-autre :

- **formes idiomatiques :** de plus en plus, par dessus tout, toute chose égale par ailleurs, etc. dont la traduction est fixe.
- **formes sémantiques limitées :** identifiant des unités comme hydraulic break pour éviter par exemple d'interpréter de manière erronée hydraulic break fluide ; ou encore des formes composées comme pomme de terre/potato
- **informations homographiques :** dressant le contexte syntaxique nécessaire à la résolution de formes homographes.
- **exception aux règles syntaxiques :** comme par exemple nor en anglais qui fait exception à la règle des autres conjonctions en permettant l'inversion du groupe verbal qui suit : ... nor could he see the difficulties
- **informations sémantiques contextuelles :** pour le choix lexical. Ex. grow $\rightarrow$ grandir (dictionnaire principal), mais peut être traduit par élever si le complément est "animé", ou encore cultiver si le complément est une plante

Systran de l'intérieur (pipeline)

Étapes de pré-traitement :

1. Identification de formatage (titres, paragraphes, indentation, etc)
2. Repérage des formes idiomatiques
3. Consultation du dictionnaire principal
4. Analyse morphologique (sauf si la langue source est l'anglais). Inférence éventuelle dans le cas de mots inconnus
5. Identification des noms composés (par consultation des dictionnaires sémantiques limités).

   Prise de décision dans des cas comme : Il parla à la femme de ménage (charlady ?).

rali

Systran de l'intérieur

Étapes d'analyse :

6. Résolution d'homographes (par analyse du contexte). Ex : states peut être un verbe ou un nom, mais pas lorsque précédé de many.
7. Segmentation des phrases en clauses (marqueurs = ponctuations, conjonctions de coordination, pronoms, etc).
8. Identification de relations syntaxiques simples (entre noms et adjectifs, temps de la phrase, son mode, etc.)
9. Désambiguisation des énumérations :

   Smog and pollution control are important factors
   Smog and pollution control is under consideration
10. Identification sujet/prédicats (traitement simpliste)
11. Identification de relations syntaxiques "profondes"

rali

Systran de l'intérieur

Étapes de Transfert :

12 Transfert des idiomes conditionnels. Ex : agree sera traduit par convenir s'il est à la forme passive et par être d'accord sinon.

13 Traduction des prépositions

14 Transfert structurel (par consultation de règles spécifiques) :
Ex : He **expects** to come / Il s'attend à ce qu'il vienne

Étapes de génération :

15 Affectation de la traduction par défaut aux mots encore à traduire

16 Génération morphologique

17 Arrangements finaux : perturbation de l'ordre des mots, élisions (le homme $\rightarrow$ l'homme).

rali

Appariement automatique de textes

Définition : tâche consistant à mettre en correspondance dans un corpus bilingue les phrases qui sont en relation de traduction.

Intérêt : mémoires de traduction, extraction de lexiques bilingues et de dictionnaires terminologiques, désambiguisation sémantique, etc.

- aligner des mots $\Rightarrow$ aligner des phrases $\Rightarrow$ aligner des mots
- des indices plutôt simples permettent d'aligner de manière "satisfaisante" des corpus au niveau de la phrase.
- en utilisant de l'information métrique et de l'information à caractère "linguistique".

[Brown et al., 1991, Debili, 1992, Gale and Church, 1993a, Simard et al., 1992, Kay and Röscheisen, 1993, Fung and Church, 1994, Wu, 1994, Langé and Gaussier, 1995, Simard and Plamondon, 1996, Melamed, 1997, Chang and Chen, 1997, Moore, 2002, Li et al., 2010]

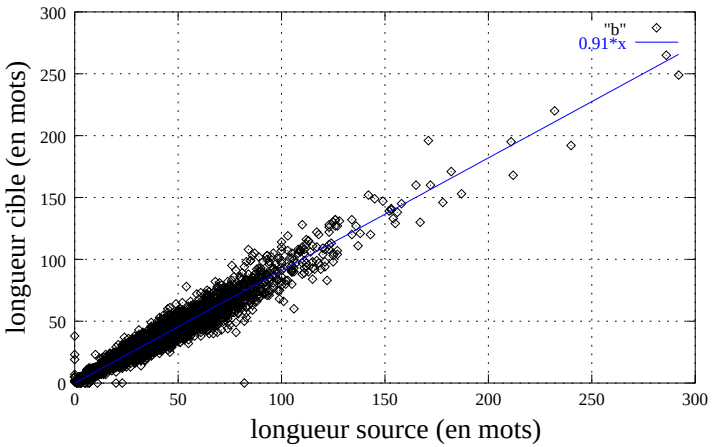
rali

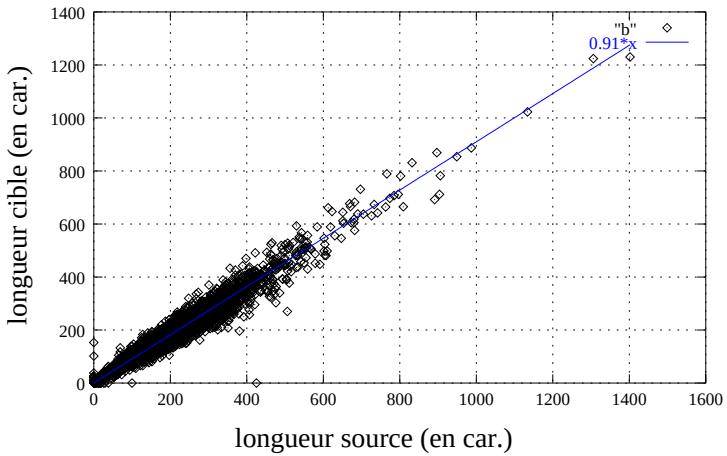
Appariement automatique de textes

Débat L'intelligence artificielle	Artificial intelligence A Debat
Depuis 35 ans, les spécialistes d'intelligence artificielle cherchent à construire des machines pensantes. Leurs avancées et leurs insuccès alternent curieusement.	Attempts to produce thinking machines have met during the past 35 years with a curious mix of progress and failure.
	Two further points are important.
Les symboles et les programmes sont des notions purement abstraites.	First, symbols and programs are purely abstract notions.

Indices métriques (longueurs)

- Le processus de traduction tend à préserver la longueur (comptée en mots ou en caractères) des phrases : il est assez probable qu'une phrase courte soit traduite par une phrase courte, et qu'à l'inverse une phrase longue soit traduite par une phrase longue.
- Cette idée simple a été exploitée en premier lieu par ***[Brown et al., 1991, Gale and Church, 1993a]***.





Corpus ilo - 7124 paires de phrases

Indices métriques (patterns)

- En général, à une phrase source est alignée une phrase cible. De manière moins fréquente, n phrases sources sont alignées avec m phrases cibles.

Corpus ilo : 7124 paires de phrases					
pattern	nb	pattern	nb	pattern	nb
1/1	6744 (94.5%)	4/1	15	1/3	3
2/1	203 (2.85%)	1/0	13	5/1	3
1/2	48 (0.6%)	0/1	6	0/9	1
3/1	37 (0.5%)	0/2	5	1/4	1
2/2	26 (0.3%)	2/0	5	1/14	1

- ***[Gale and Church, 1993a]*** propose les six patrons suivants : 1/1 (0.89), 1/2 et 2/1 (0.089), 0/1 et 1/0 (0.011) et 2/2 (0.010)

Indices métriques

- ***[Gale and Church, 1993a]*** mesure la qualité d'un appariement de l_1 caractères sources avec l_2 caractères cibles par $p(match|\delta)$:

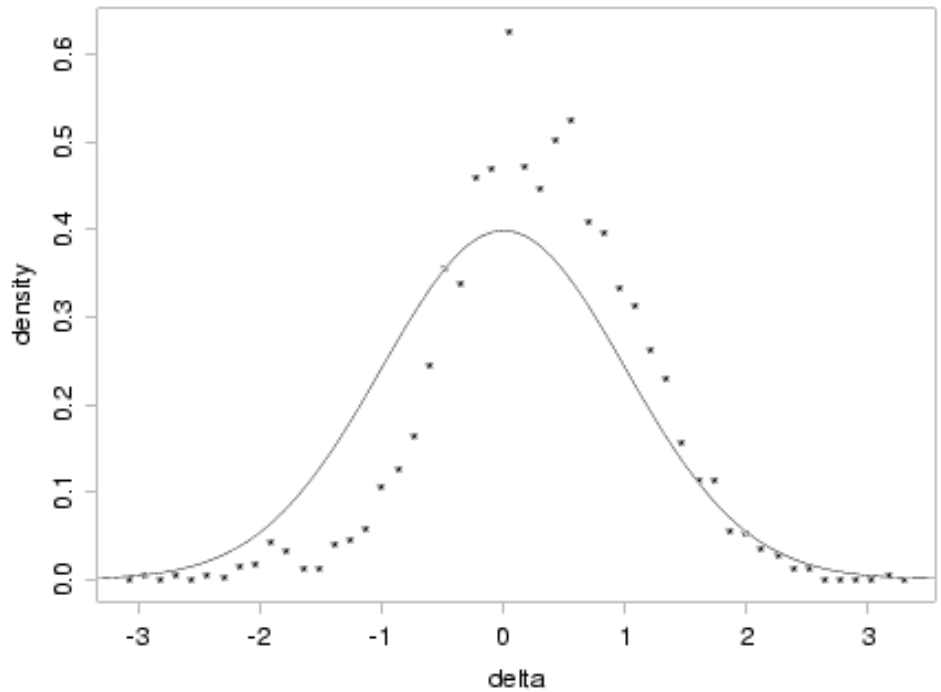
$$\begin{aligned} S_{gc} &= -\log(p(\delta|match) \times p(match)) \\ p(\delta|match) &= 2 \times (1 - p(|\delta|)) \end{aligned}$$

où δ suit une loi normale centrée réduite : $\delta = \frac{(l_2 - l_1 c)}{\sqrt{l_1 s^2}}$ qui dépend de deux paramètres c (la moyenne) et s^2 (la variance) déterminés expérimentalement (resp. 1 et 6.8).

- $p(match)$ voir acétate d'avant

Indices métriques

- **Hypothèse :** Chaque caractère de L_1 donne lieu à un nombre aléatoire de caractères dans L_2. Le nombre de caractères dans L_2 "généré" par un caractère de L_1 est une variable aléatoire de moyenne c et de variance s^2 :



En pointillé : distribution empirique des δ. En trait plein, une normale centrée réduite.

Indices à caractère linguistique

- *[Simard et al., 1992]* propose d'exploiter le fait que deux phrases en relation de traduction partagent souvent des mots communs ou proches (chiffres, noms propres, etc).
- Deux mots sont dits cognates s'ils partagent des propriétés communes à un quelconque niveau (sémantique, orthographique, etc.).

  accès/access, activité/activity, parlement/parliament
  librairie/library
- **Définition** *[Simard et al., 1992]* **:** deux mots sont cognates s'ils ont en commun un **préfixe de 4 caractères** au moins. Si l'un des mots contient un chiffre au moins, alors les deux mots sont cognates s'ils sont égaux.
- Voir *[Ribeiro et al., 2001]* pour un algorithme reconnaissant plus de cognates (ex : gouvernement/government).

rali

Indices à caractère linguistique

- ***[Simard et al., 1992]*** qualité d'un appariement : $S_{cog} = \frac{P_T(c|n)}{P_R(c|n)} \times p(match)$ où $P_T(c|n)$ et $P_R(c|n)$ dénotent la probabilité que deux segments de longueur moyenne n (en mots) possèdent c cognates sous les hypothèses respectives que les segments sont en relation de traduction, ou au contraire qu'ils ont été sélectionnés aléatoirement.
- Les auteurs observent sur une partie étiquetée du Hansard, que ces deux probabilités suivent approximativement des **lois binomiales** où p_T (resp. p_R) désigne la probabilité qu'un mot d'un segment ait un cognate dans un segment de l'autre langue lorsque ces deux segments sont (resp. ne sont pas) traduction l'un de l'autre (p_T et p_R ont été expérimentalement fixés à 0.3 et 0.09).

$$\begin{aligned} p_T(c|n) &= \binom{n}{c} \times p_T^c \times (1-p_T)^{n-c} \\ p_R(c|n) &= \binom{n}{c} \times p_R^c \times (1-p_R)^{n-c} \end{aligned}$$

Algorithme d'alignement

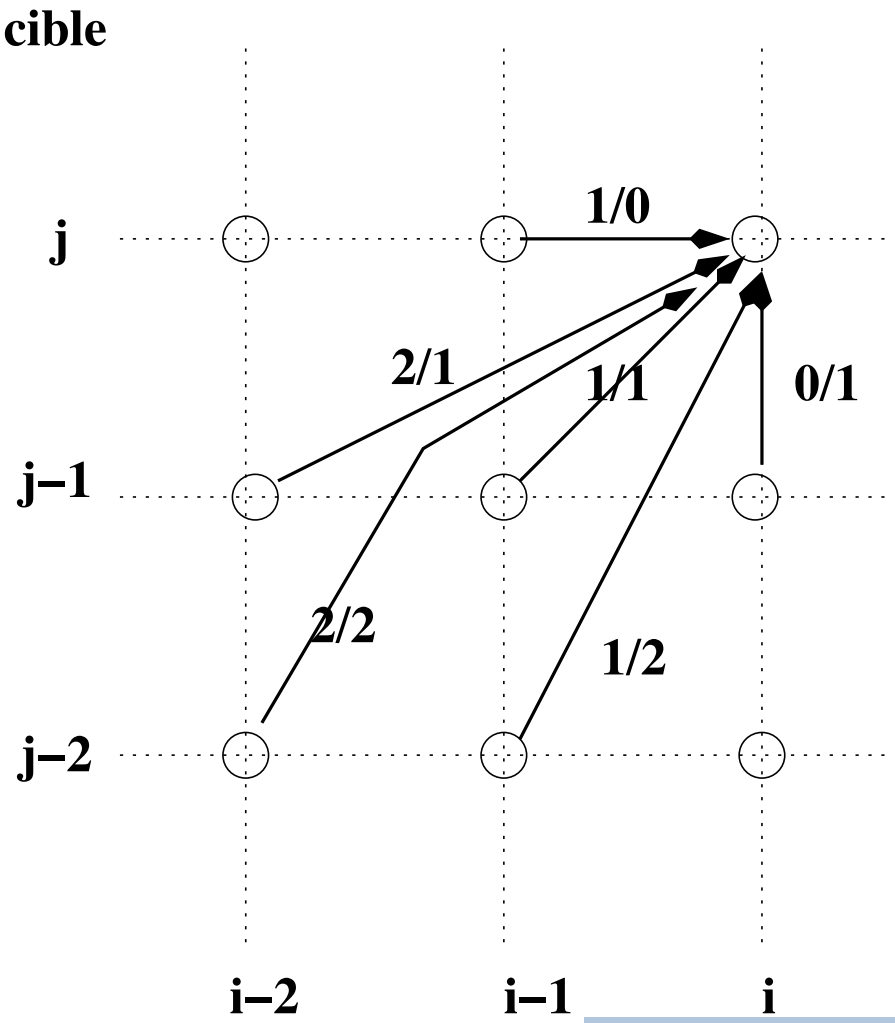
- Par programmation dynamique ; soit, $D(i,j)$ le (score du) meilleur alignement des phrases sources $s_1, \ldots, s_i$ avec les phrases cibles $t_1, \ldots t_j$.
  - À l'initiale, $D(i,j)$ est nul pour tout i et j
  - Récurrence, sur les schémas d'appariement : 0/1, 1/0, 1/1, 1/2, 2/1 et 2/2 :

$$D(i,j) = min \begin{cases} D(i,j-1) & + & d_1 \\ D(i-1,j) & + & d_2 \\ D(i-1,j-1) & + & d_3 \\ D(i-1,j-2) & + & d_4 \\ D(i-2,j-1) & + & d_5 \\ D(i-2,j-2) & + & d_6 \end{cases}$$

où $d_{i \in [1,6]}$ sont les coûts respectifs de chaque schéma selon un score donné (par exemple S_{gc} ou S_{cog}).

Algorithme d'alignement

Comme dans viterbi, on garde un pointeur de retour arrière sur la meilleure case de laquelle on vient.



Algorithme d'alignement : codage

- Nous avons besoin de deux tables consignant les informations de chaque phrase des textes source et cible.

```
INFO Source[nbPhraseSource];
INFO Cible[nbPhraseCible];
```

où `INFO` contient toutes les informations nécessaires au calcul du score d'un appariement.

Ex :

$$\text{INFO} = \left[\begin{array}{ll} \text{int carLength}; & \text{//longueur de la phrase en caractères} \\ \text{int wordLength}; & \text{//longueur de la phrase en mots} \\ \text{string texte}; & \text{//la phrase elle-même (pas nécessaire)} \end{array}\right.$$

rali

Algorithme d'alignement : codage

- Nous avons également besoin d'une représentation de l'espace de recherche (une matrice) :

```
DTW Space[nbPhraseSource][nbPhraseCible]
```

$$\text{DTW} = \left[\begin{array}{ll} \text{float score}; & \text{//le (meilleur) score} \\ \text{pair}\langle int, int \rangle \text{ back}; & \text{//le "pointeur" de retour (un entier suffit)} \end{array} \right.$$

- **Note :** Aligner un corpus bilingue de 1000 phrases sources avec 1000 phrases cibles requiert :
  - $1000 \times 1000 \times (8 + 2 \times 4)$ soit $16\,000\,000$ octets (16 Meg !).

$\rightarrow$ réduction de l'espace de recherche + matrice creuse

Algorithme d'alignement : codage

- Le score de l'alignement des di phrases sources ($S[i] \dots S[i - di + 1]$) avec les dj phrases cibles ($T[j] \dots T[j - dj + 1]$) est donnée par :

```
float score(int i, int j, int di, int dj, ...)
```

- le score d'un alignement est additif (somme des scores des appariements = score de l'alignement).
- Lorsque possible, il est souhaitable de précalculer des informations (ex : les mots cognates)

Algorithme d'alignement : codage

Space[i][j].score $\leftarrow 0 \; \forall i \in [1, nbPhSource], j \in [1, nbPhCible]$
for (int i=0 ; i<nbPhSource ; i++) **do**
  path $\leftarrow$ false
  for (int j=0 ; i<nbPhCible ; j++) **do**
    if (inSpace(i,j)) **then**
      // alignment 1/1
      if (inSpace(i-1,j-1)) **then**
        score $\leftarrow= Score(i, j, 1, 1)$
        if (score $>$ Space[i][j].score) **then**
          Space[i][j].score = score ;
          Space[i][j].back = make_pair(1,1) ; path $\leftarrow$ true
      // alignment 2/1
      if (inSpace(i-2,j-1)) **then**
        score $\leftarrow= Score(i, j, 2, 1)$
        if (score $>$ Space[i][j].score) **then**
          Space[i][j].score = score ;
          Space[i][j].back = make_pair(2,1) ; path $\leftarrow$ true
      // on passe de même tous les alignements possibles
  if (path == false) **then**
    // Alignement impossible (impasse sur la phrase i)

Algorithme d'alignement : retour du meilleur chemin

- Le meilleur alignement a pour score :
  `Space[nbPhSource][nbPhCible].score`
- et l'alignement est obtenu en retournant
  `Space[nbPhSource][nbPhCible].back`

```
i ← nbPhSource
j ← nbPhCible
while (i && j) do
    PRINT Alignement en (i,j) avec un schéma :
    Space[i][j].back.first/Space[i][j].back.second
    i -= Space[i][j].back.first
    j -= Space[i][j].back.second
```

Réduction de l'espace de recherche

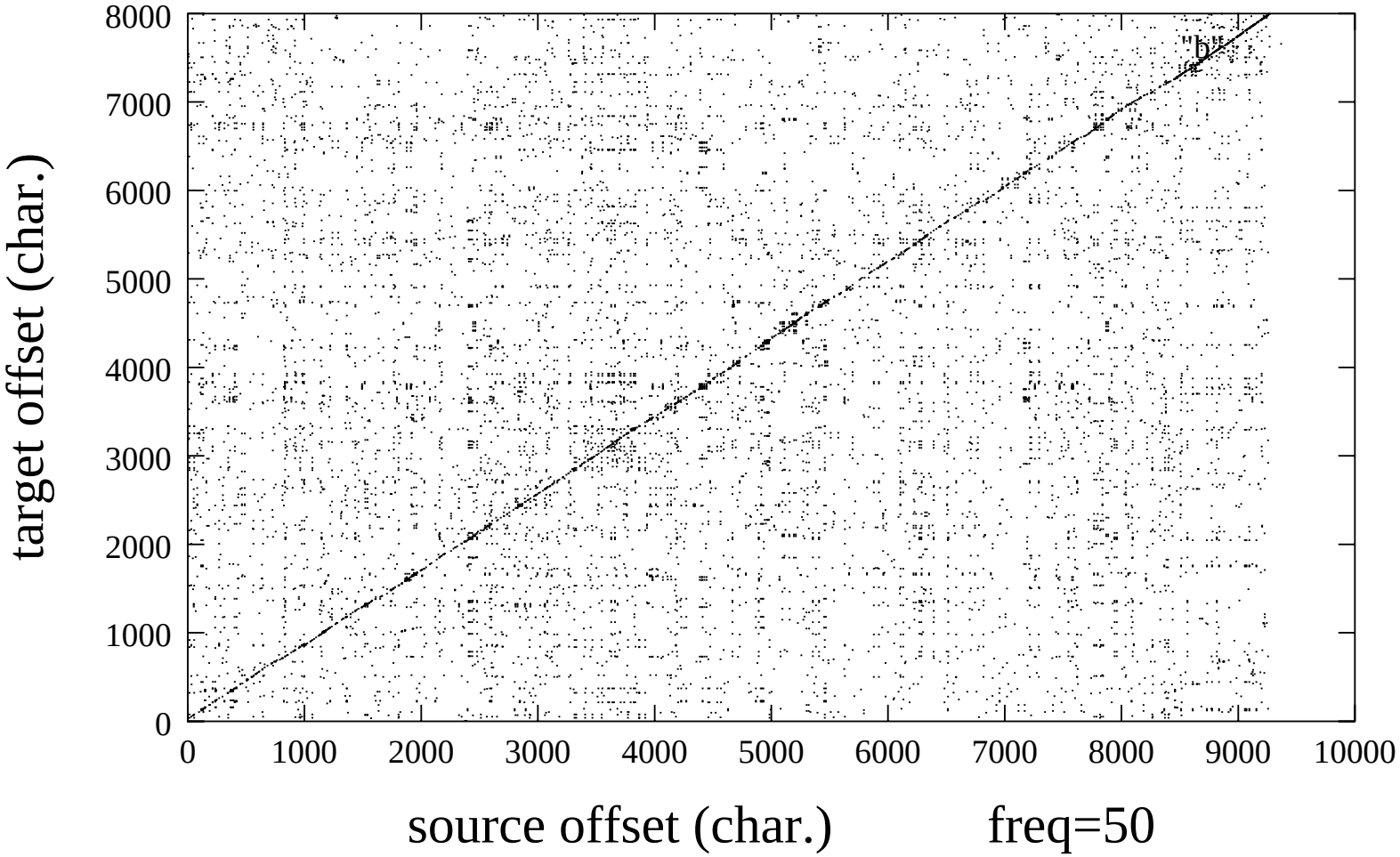
- Réduire les temps de calculs de manière significative (possiblement avec perte en robustesse)
  - **Idée :** l'alignement grossier au niveau des mots permet de réduire l'espace de recherche.
  - **hypothèse :** les cognates peu fréquents sont probablement de bons indicateurs de synchronisation des deux textes.
- Soit $M(i,j)$ une matrice qui vaut 1 si le i-ème mot source et le j-ième mot cible sont cognates, et 0 sinon.
  - on cherche à minimiser la déviation à la diagonale que l'on devrait observer si notre hypothèse est vérifiée.

$$S(I,J) = \min_{i=I-R}^{I-1} \min_{j/M_{i,j}=1} (S(i,j) + F(i,j,I,J))$$

$$\text{with } F(i,j,I,J) = \frac{J-j}{I-i} + (I-i-1) \times \mathcal{C}$$

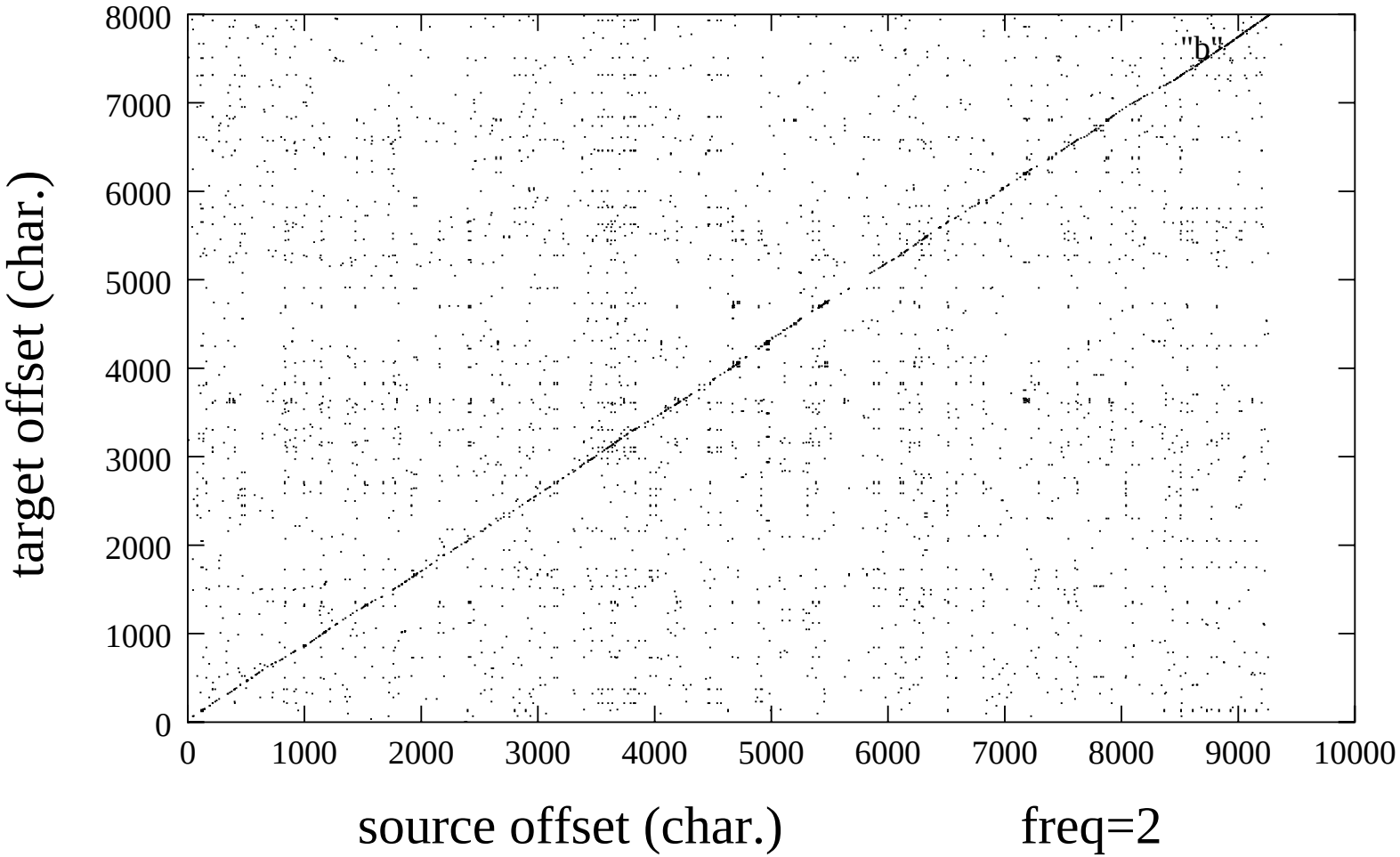
$\mathcal{C}$ et R constantes déterminées empiriquement.

Réduction de l'espace de recherche



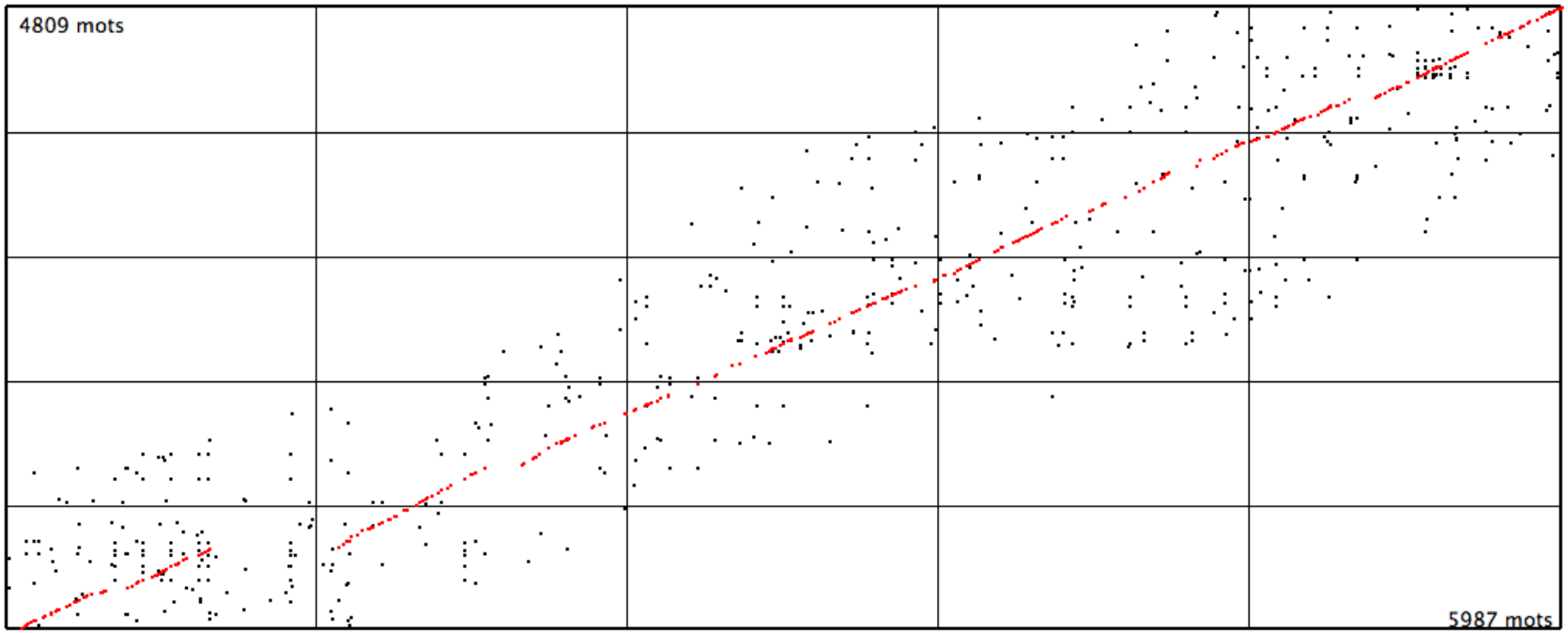
Un point correspond à deux mots qui sont cognates dans un bitexte.

Réduction de l'espace de recherche



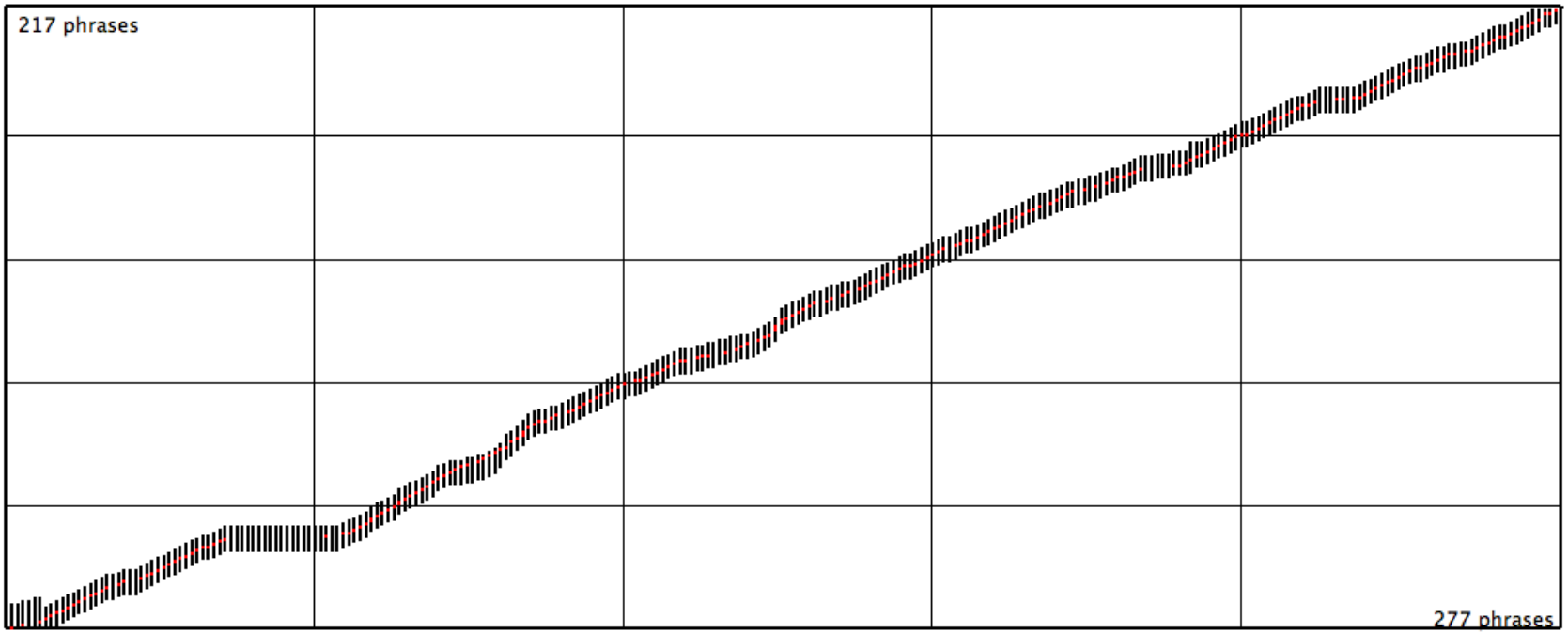
Un point correspond à deux mots qui sont cognates dans un bitexte.

Réduction de l'espace de recherche



- en rouge : le résultat de l'alignment basé sur les cognates

Réduction de l'espace de recherche



- cet alignment permet de définir un faisceau étroit sur lequel appliquer la prog. dynamique

Alignement de phrases : Évaluation

L'AUPELF-UREF finance ARCADE une action de recherche sur l'évaluation d'algorithmes d'alignement bilingues ***[Langlais et al., 1998]***.

- Une première campagne (95-96) avait pour but de collecter des corpus de référence (des bitextes alignés manuellement au niveau des phrases) et de mettre au point une méthodologie d'évaluation de la performance des algorithmes d'alignement. Cinq équipes participent aux tests :
  - RALI (Montréal), LORIA (Nancy), IRMC(Paris), ISSCO (Genève), LIA (Avignon).
- Une seconde campagne (97-98) réunit 9 équipes et étend les tests à l'alignement de mots ***[Véronis and Langlais, 2000]*** :
  + LILLA (Nice), UPenn (Pennsylvanie), CEA (Paris), XEROX (Grenoble)

Arcade : corpus de référence

BAF : $400\,000$ mots par langue (anglais/français). Quatre types de textes :

- INST ($300\,000$ mots par langue de textes institutionnels, ex : Hansard),
- ISCIENCE ($50\,000$ mots par langue d'articles scientifiques),
- ITECH (documentation technique de Xerox, $39\,328$ mots anglais, $46\,828$ mots français),
- IVERNE ("De la terre à la lune" de Jules Verne, 40161 mots anglais, $53\,181$ mots français).

JOC : $200\,000$ mots par langue du journal officiel de la communauté européenne (français/anglais). Série de questions/réponses sur différents problèmes concernant la communauté européenne. Année 93.

Arcade : Métriques d'évaluation

- Taux de précision et rappel, mais à différents degrés de granularité.
- Soit $S = \{s_1, s_2, \ldots, s_n\}$ et $T = \{t_1, t_2, \ldots, t_m\}$ sa traduction. Un alignement A entre S et T est un sous-ensemble du produit cartésien $\wp(S) \times \wp(T)$, où $\wp(S)$ et $\wp(T)$ sont respectivement l'ensemble de tous les sous-ensembles de S et T. On appelle le triplet $\langle S, T, A \rangle$ un *bitexte*.
- Voici un bitexte de référence que nous utilisons par la suite :

s_1 Ceci est la phrase numéro un.	t_1 This is the first sentence.
s_2 Ceci est la phrase numéro deux qui ressemble à la première	t_2 This is the second sentence. t_3 It looks like the first.

- La représentation de cet alignement est :

$$A_r = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2, t_3\})\}$$

Arcade : Métriques d'évaluation

- Soit le bitexte de référence $\langle S, T, A_r \rangle$ et un alignement candidat A.

$$\text{rappel} = |A \cap A_r|/|A_r| \quad \text{précision} = |A \cap A_r|/|A|$$

s_1 Ceci est la phrase numéro un.	t_1 This is the first sentence.
	t_2 This is the second sentence.
s_2 Ceci est la phrase numéro deux, qui ressemble à la première.	t_3 It looks like the first.

$$\begin{aligned} A &= \{(\{s_1\},\{t_1\}),(\{\},\{t_2\}),(\{s_2\},\{t_3\})\} \\ A_r &= \{(\{s_1\},\{t_1\}),(\{s_2\},\{t_2,t_3\})\} \end{aligned}$$

- $A \cap A_r = \{(\{s_1\},\{t_1\})\} \Rightarrow$ rappel=1/2, précision=1/3 et F-mesure = $2 \times \frac{p \times r}{p+r}$ = 0.4 (*métriques MA dans la table*)

Arcade : Métriques d'évaluation

- Les taux de précision et rappel tels que calculés précédemment sont sévères car des *bisegments* (éléments de l'alignement) sont peut-être partiellement corrects.
  - Dans l'exemple précédent, $(\{s_2\}, \{t_3\})$ n'appartient pas à A_r mais t_3 est bel et bien associé à s_2 dans la référence.

$\rightarrow$ donner du crédit à des alignments partiellement corrects en considérant le produit cartésien des alignements dans chaque bisegment

- Posons $A = \{a_1, a_2, \ldots, a_m\}$ et $A_r = \{ar_1, ar_2, \ldots, ar_n\}$ avec $a_i = (as_i, at_i)$ et $ar_j = (ars_j, art_j)$, alors :

$$\begin{array}{lcll} A' & = & \bigcup_i (as_i \times at_i) & = & \{(s1,t1),(s2,t3)\} \\ A'_r & = & \bigcup_j (ars_j \times art_j) & = & \{(s1,t1),(s2,t2),(s2,t3)\} \end{array}$$

- précision=1 et rappel=2/3, F-mesure=0.8 (au lieu de 0.4)
  (métriques MP dans la table)

Arcade : Métriques d'évaluation

Problèmes avec ces métriques calculées au niveau de la phrase :

1. Nécessite une segmentation au niveau de la phrase
2. Tous les algorithmes n'en font pas usage
3. Ne donne pas d'information sur la gravité d'un alignement (une erreur sur une petite phrase est peut-être moins pénalisant dans une application, qu'une erreur commise sur un long passage).

$\Rightarrow$ *mesures des taux de précision et rappel au niveau des caractères (ou des mots)*

rali

Arcade : Métriques d'évaluation

- Sur notre exemple, au niveau des mots :

```
|Ar'| = 5*6 + 11*10 = 140
|A'| = 5*6 + 0*5 + 11*6 = 96
|Ar' ^ A'|  = 96
recall = 96/140 = 0.69
precision = 1
F = 0.82
```

- Sur notre exemple, au niveau des caractères (incluant les espaces) $\hookrightarrow$ *métriques* MC *dans la table*

```
|Ar| = 29*27 + 60*52 = 3903
|A| = 29*27 + 0*28 + 60*24 = 2223
|Ar' ^ A'| = 2223
recall = 2223/3903 = 0.57
precision = 1
F = 0.73
```

Arcade : Performances (Corpus BAF)

	MP			MC			MA		
	Pr	Rc	F	Pr	Rc	F	Pr	Rc	F
APA	97.01	81.24	88.43	98.19	88.92	93.33	87.95	91.72	89.80
GSA+	95.24	81.57	87.88	97.18	88.97	92.89	88.78	90.59	89.68
GSA	94.98	81.08	87.48	97.38	88.65	92.81	88.33	90.36	89.34
SFI	94.19	81.66	87.48	95.96	88.55	92.11	88.40	89.23	88.81
LILLA	92.97	78.79	85.29	96.06	86.26	90.90	84.45	86.59	85.51
JACAL	91.66	79.73	85.28	94.06	87.15	90.47	82.26	86.63	84.39
SALIGN	85.42	82.98	84.18	93.12	89.26	91.15	88.85	85.35	87.06
GC	90.60	77.39	83.47	92.82	84.65	88.55	83.96	85.29	84.62
ISSCO	88.95	76.33	82.16	91.00	82.25	86.41	83.31	84.19	83.75
CEA	92.56	71.08	80.41	98.10	83.16	90.01	77.91	82.50	80.14
LORIA	86.41	72.65	78.94	90.80	80.88	85.56	80.26	82.27	81.25
IRMC	81.22	65.74	72.67	83.38	73.38	78.06	70.86	76.85	73.74

Arcade : Performances (Corpus JOC)

	MP			MC			MA		
	Pr	Rc	F	Pr	Rc	F	Pr	Rc	F
GSA+	99.08	98.83	98.95	99.58	99.41	99.49	98.14	98.40	98.27
APA	99.23	98.64	98.93	99.63	98.93	99.28	97.82	98.41	98.11
LORIA	98.89	98.82	98.86	99.50	99.01	99.26	98.41	98.37	98.39
GSA	98.86	98.71	98.78	99.51	99.34	99.42	97.96	98.14	98.05
SFI	98.33	99.02	98.67	99.03	99.26	99.14	98.41	98.20	98.31
JACAL	97.93	98.67	98.30	98.78	99.36	99.07	97.91	97.54	97.72
GC	96.94	97.17	97.06	98.10	97.97	98.03	96.52	96.33	96.43
LILLA	97.19	96.93	97.06	98.39	96.85	97.61	95.51	95.66	95.59
IRMC	97.10	96.92	97.01	98.37	98.31	98.34	93.84	94.90	94.37
SALIGN	94.73	99.20	96.92	97.82	99.51	98.66	96.99	95.04	96.01
ISSCO	95.02	95.80	95.41	96.82	96.67	96.74	95.74	95.06	95.40
CEA	96.58	91.05	93.74	98.38	93.18	95.71	88.43	90.45	89.43

rali

À propos des aligneurs de phrases

- Il en existe plusieurs en code ouvert :

  G&C *[Gale and Church, 1993b]* (le code est dans l'article !)
  BMA *[Moore, 2002]*
  HUNALIGN *[Varga et al., 2005]*
  YASA `http://rali.iro.umontreal.ca/rali/?q=en/yasa`

- plusieurs études montrent qu'on peut itérativement sélectionner des alignments de phrases fiables de manière à :
  - entrainer un classificateur à reconnaître des paires de phrases parallèles *[Yu et al., 2012]*
  - entraîner un système de traduction pour traduire les phrases et faire une comparaison monolingue *[Sennrich and Volk, 2011]* ou pour regrouper des alignments 1-n ou n-1 *[Braune and Fraser, 2010]*
- d'autres s'affranchissent partiellement ou en totalité de l'hypothèse de monotonicité
  [Bisson and Fluhr, 2000, Deng et al., 2007, Quan et al., 2013]

rali

BMA, YASA et HUNALIGN sont dans un bateau

bitext	YASA		BMA		HUNALIGN	
	F_A	F_S	F_A	F_S	F_A	F_S
INST	**94.9**	**96.4**	93.3	91.4	91.0	93.4
SCIENCE	**89.4**	**91.7**	88.9	86.4	84.8	86.5
VERNE	69.2	**86.9**	**72.3**	74.6	66.0	74.6
TECH	90.4	**10.9**	**94.2**	10.3	89.6	10.7

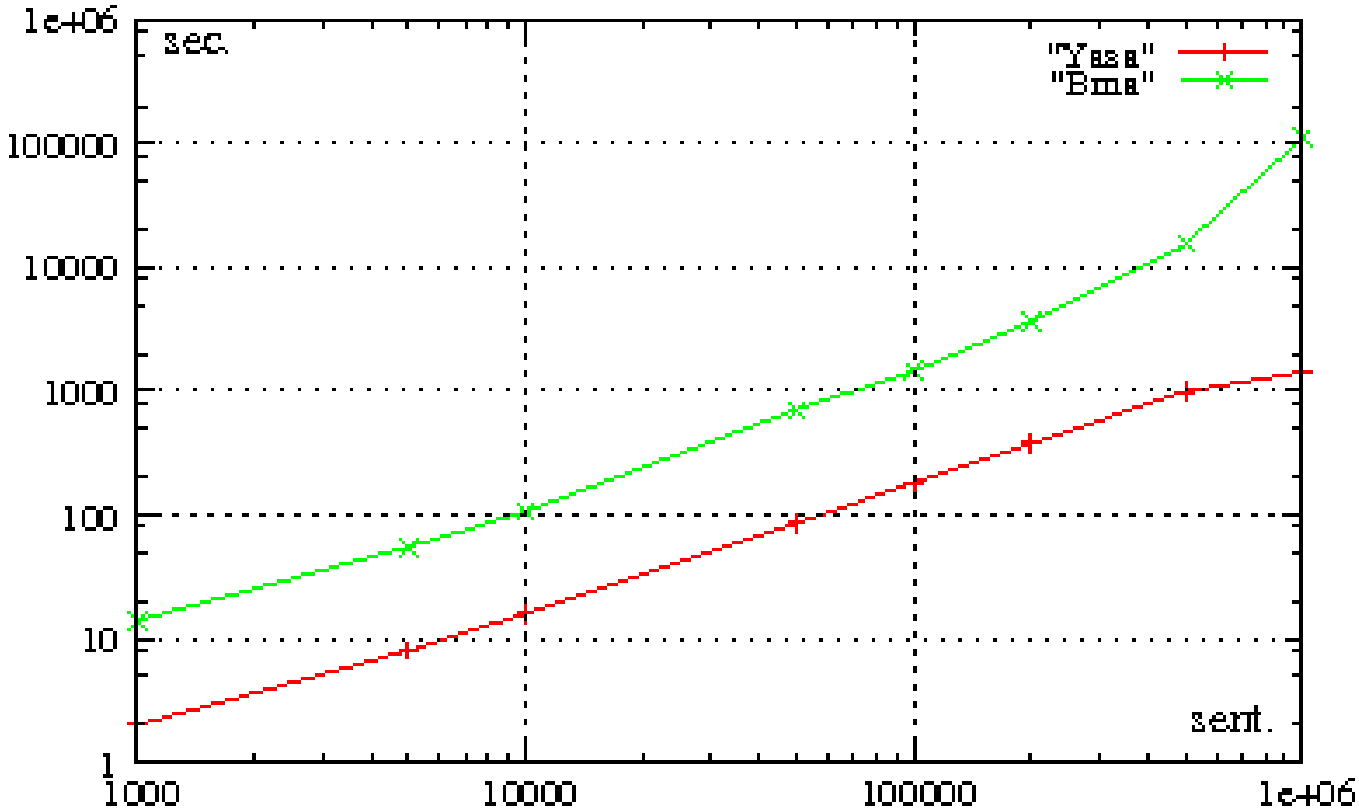
- Résultats sur BAF (F_A : niveau alignement, F_S : niveau phrase)

BMA, YASA et HUNALIGN sont dans un bateau

%	INST	SCIENCE	VERNE	TECH
0	94.9 (+1.6)	89.4 (+0.5)	69.2 (-3.1)	90.4 (-3.8)
10	85.9 (+1.1)	78.8 (-1.9)	56.1 (-2.5)	72.2 (-7.1)
20	81.7 (+2.2)	75.2 (+2.9)	54.5 (-3.8)	74.7 (-6.5)
30	77.8 (+7.0)	70.5 (+8.3)	36.7 (-7.6)	69.1 (-3.8)
40	76.0 (+16.6)	69.1 (+30.5)	40.6 (+3.3)	65.0 (+1.8)
50	69.0 (+23.0)	67.2 (+41.8)	44.7 (+28.8)	64.3 (+14.9)
60	69.5 (+44.8)	67.8 (†)	49.9 (+41.9)	65.8 (†)

- Introduction artificielle de bruit (en supprimant des phrases dans la source)
- Entre parenthèse : gain absolu par rapport à BMA

BMA, YASA et HUNALIGN sont dans un bateau



YASA & Traduction — en-fr

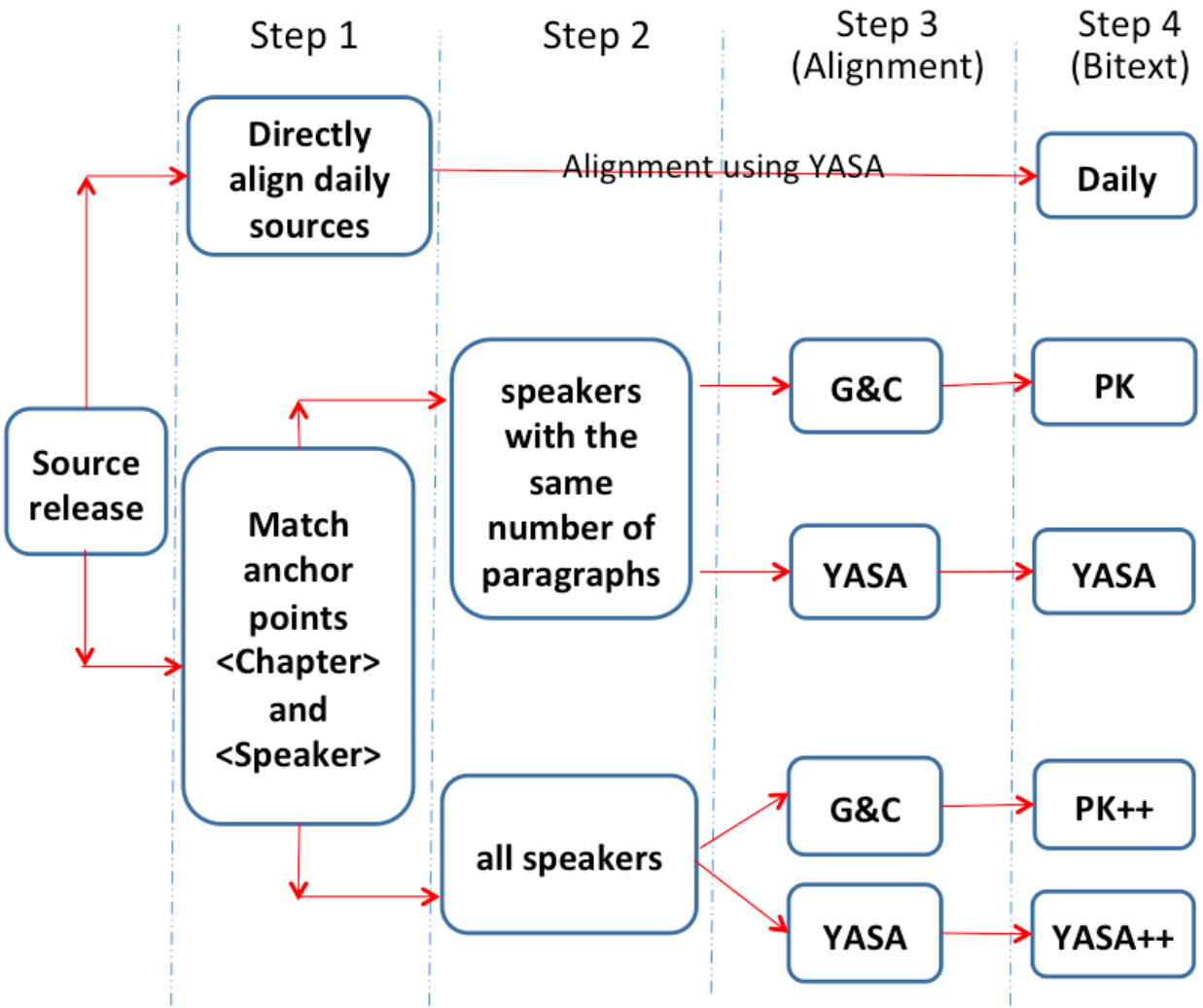
- EUROPARL version pré-alignée :
  `http://www.statmt.org/europarl/`

<CHAPTER ID=1>	<CHAPTER ID=1>
<SPEAKER ID=1 NAME="La présidente">	<SPEAKER ID=1 NAME="President">
Reprise de la session	Resumption of the session
<P>	<P>
Je déclare reprise de la session du parlement européen qui avant été interrompue le Vendredi 17 Décembre. . .	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999 . . .
<P>	<P>

- `sentence-align-corpus.perl`

1. CHAPTER supposés synchronisés (ID est ignoré)
2. SPEAKER supposés synchronisés (ceux en trop sont ignorés)
3. les tours de paroles où le nombre de P est identique sont alignés à l'aide de G&C.

YASA & Traduction — en-fr



Qualité d'un bitexte versus performance en traduction

	NEWS09	NEWS10	NEWS11	HANS
DAILY	**20.02**	20.46	**21.16**	23.64
PK	19.43	20.14	20.77	23.46
YASA	19.81	20.47	20.93	23.60
PK++	19.49	20.27	20.62	23.25
YASA++	19.96	**20.61**	21.05	**23.95**

- Performance (BLEU) sur différents test sets (fr→en)
- gains également sur d'autres paires de langues (***[Lamraoui and Langlais, 2013]***)
- voir aussi ***[Goutte et al., 2012]*** qui montre (au contraire) que du bruit dans les bitextes affectent peu les performances BLEU

rali

Quelques perles

Extraites de http://www.languagetranslation.com/resources/humorous-examples.html

- Cocktail lounge, Norway :

  Ladies are Requested Not to have Children in the Bar
- At a Budapest zoo :

  PLEASE DO NOT FEED THE ANIMALS. If you have any suitable food, give it to the guard on duty
- Hotel, Acapulco :

  The Manager has Personally Passed All the Water Served Here
- Tokyo hotel's rules and regulations :

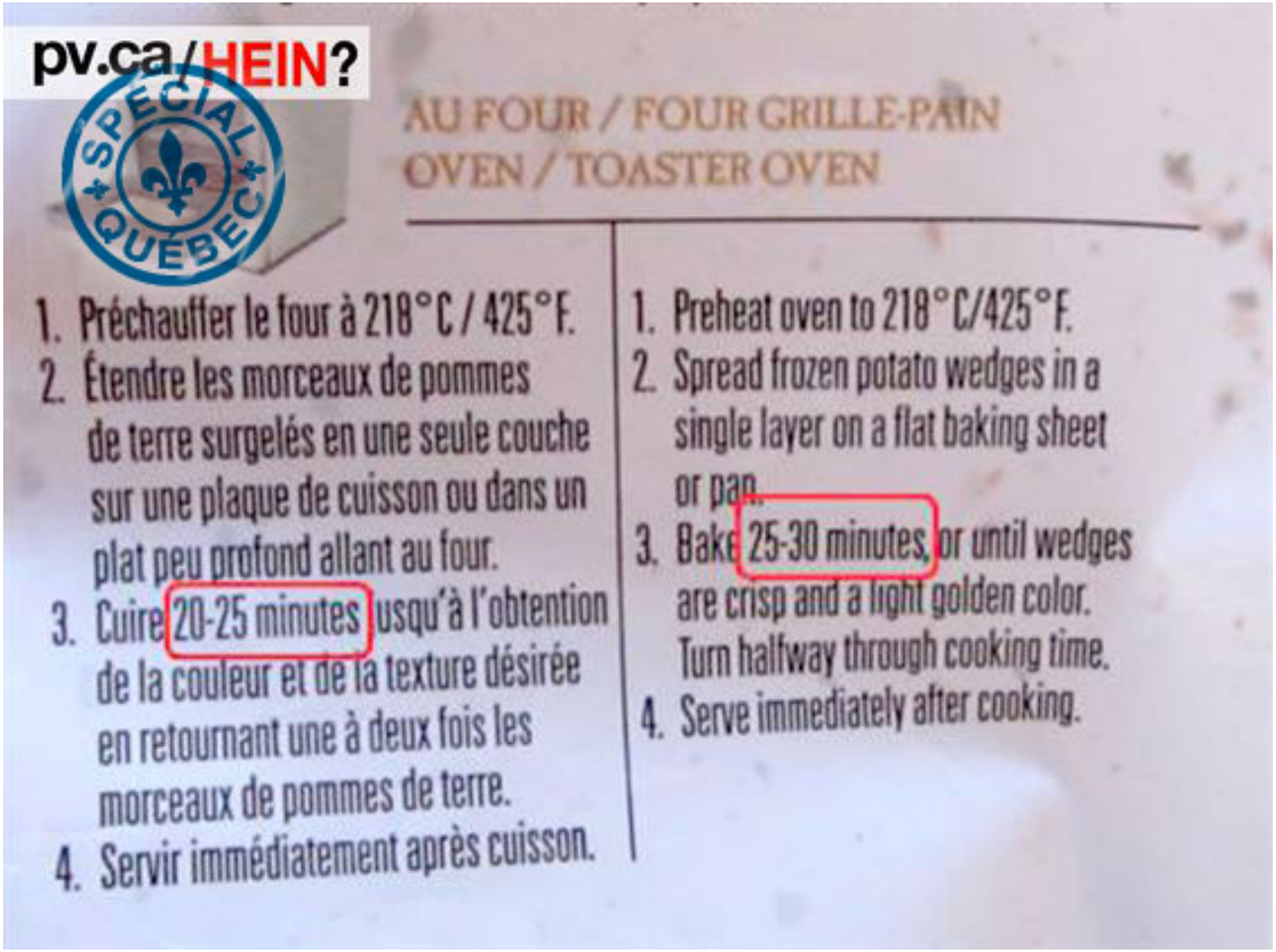
  Guests are requested NOT to smoke or do other disgusting behaviors in bed.
- On the menu of a Swiss restaurant :

  Our wines leave you nothing to hope for.
- Hotel elevator, Paris :

  Please leave your values at the front desk.
- Hotel, Japan :

  You are invited to take advantage of the chambermaid.

Quelques perles



Quelques perles

BABELFISH, GOOGLE et SYSTRAN sont dans un bateau 1/3

b : http://www.babelfish.fr/
s : http://www.systran.fr/lp/traduction-en-ligne/
g : https://translate.google.ca/?hl=fr (before 2015) G : https://translate.google.ca/?hl=fr 2018

He likes her

b	Il aime lui
s,g,G	Il l'aime

She likes him

b	Elle aime lui
s,g,G	Elle l'aime

Il l'aime

b	He loves her
s	He likes it
g	He loves
G	He loves it

Elle l'aime

b,G	She loves him
s	She likes it
g	She loves

BABELFISH, GOOGLE et SYSTRAN sont dans un bateau 2/3

He took a french leave

g,G	il a pris un congé Français
b	Il a pris un congé de Français
s	Il a pris le congé de Français

elle donne du fil à retordre à ses parents

b	It gives of trouble to his parents
s	it gives wire to be twisted with her parents
g	it gives a headache to his parents
G	she gives her parents a hard time

BABELFISH, GOOGLE et SYSTRAN sont dans un bateau 3/3

you should never trust machine translations

g	vous ne devriez jamais faire confiance traductions automatiques
b	vous ne devriez jamais confiance traductions automatiques
s	vous devriez ne jamais faire confiance à des traductions automatiques
G	vous ne devriez jamais faire confiance aux traductions automatiques

Paul aime virginie qui ne le regarde même pas

b	Paul loves Virginia which doesn't even look not
s	Paul likes Virginia which does not even look at it
g	Paul looks like Virginia that not even
G	Paul loves virginia who does not even look at him

rali

Exemple de sorties de WMT 2016 (de2en)

S	Obama empfängt Netanyahu
R	Obama receives Netanyahu
A	Obama, Netanyahu
B	Obama welcomes Netanyahu
C	Obama welcomes Netanyahu

S	Das Verhältnis zwischen Obama und Netanyahu ist nicht gerade freundschaftlich.
R	The relationship between Obama and Netanyahu is not exactly friendly.
A	The relationship between Obama and Netanyahu is not exactly friendly.
B	The relationship between Obama and Netanyahu is not exactly friendly.
C	The relationship between Obama and Netanyahu is not very friendly.

Exemple de sorties de WMT 2016 (de2en)

S	Die beiden wollten über die Umsetzung der internationalen Vereinbarung sowie über Teherans destabilisierende Maßnahmen im Nahen Osten sprechen
R	The two wanted to talk about the implementation of the international agreement and about Teheran's destabilising activities in the Middle East.
A	The two wanted to talk about the implementation of the international agreement on Tehran's destabilizing policies in the Middle East.
B	The two wanted to talk about the implementation of the international agreement, as well as Tehran's destabilizing measures in the Middle East.
C	The two wanted to talk about the implementation of the international agreement on Tehran's destabilizing actions in the Middle East.

Exemple de sorties de WMT 2016 (de2en)

S	Bei der Begegnung soll es aber auch um den Konflikt mit den Palästinensern und die diskutierte Zwei-Staaten-Lösung gehen.
R	The meeting was also planned to cover the conflict with the Palestinians and the disputed two state solution.
A	At the meeting, but it is also about the conflict with the Palestinians and discussed the two-state solution.
B	At the meeting, however, it is also a question of the conflict with the Palestinians and the two-state solution that is being discussed.
C	At the meeting it will go, but also about the conflict with the Palestinians and the two state solution under discussion.

Exemple de sorties de WMT 2016 (de2en)

S	Das Verhältnis zwischen Obama und Netanyahu ist seit Jahren gespannt.
R	Relations between Obama and Netanyahu have been strained for years.
A	The relationship between Obama and Netanyahu is looking forward for years.
B	The relationship between Obama and Netanyahu has been tense for years.
C	The relationship between Obama and Netanyahu has been strained for years.

Exemple de sorties de WMT 2016 (de2en)

S	Washington kritisiert den andauernden Siedlungsbau Israels und wirft Netanyahu mangelnden Willen beim Friedensprozess vor.
R	Washington criticises the continuous building of settlements in Israel and accuses Netanyahu of a lack of initiative in the peace process.
A	Washington criticized the ongoing Israeli settlement construction and accuses Netanyahu lack of faith in the peace process.
B	Washington criticizes Israel's ongoing settlement construction, accusing Netanyahu of a lack of will in the peace process.
C	Washington condemned the ongoing Israeli settlement construction, and Netanyahu blames a lack of faith in the peace process.

Exemple de sorties de WMT 2016 (de2en)

A système à base de segments

B système neuronal

C système syntaxique

Les trois systèmes sont réalisés au sein de la même équipe (UofEdin)

Fluency (fluidité) / Adequation (fidélité)

- Grille d'évaluation :

Fluency	
5	Flawless English
4	Good English
3	Non-Native English
2	Disfluent English
1	Incomprehensible

Adequacy	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

- jugements bruités : l'humain s'habitue !
- tâche difficile et longue (coûteuse)

Classement de systèmes 1/2

SRC	A good idea, provided that it will, prove realistic this time.
REF	Une bonne idée, à condition de se montrer réaliste, cette fois.
S1	Une bonne idée, à condition qu'il se montre réaliste, ceci chronomètre.
S2	Une bonne idée, à condition qu'il sera cette fois réaliste.
S3	Une bonne idée, pourvu qu'elle va s'avérer réaliste cette fois-ci.
S4	Une bonne idée, pourvu qu'il va s'avérer réaliste cette fois.
S5	Une bonne idée, à condition qu'il aille, s'avère réaliste cette fois.
S6	Une bonne idée, à condition qu'elle s'avèrera, réaliste cette fois.
S7	Une bonne idée, à condition qu'elle va se montrer réaliste, cette fois.
S8	A fourni une bonne idée, qu'il va, s'avérer réalistes pour l'instant.
S9	Une bonne idée, à condition que ce sera, cette fois prouver réaliste.
S10	Une bonne idée, pourvu que lui will, se révéler réaliste cette fois-ci.
S11	Une bonne idée, à condition qu'il aille, s'avère réaliste cette fois.
S12	Une bonne idée, fournie qu'il fait, prouve réaliste cette fois.
S13	Une bonne idée, à condition que elle, s'avérer réaliste cette fois.
S14	Une bonne idée, à condition qu'il faudra prouver réaliste cette fois.
S15	Une bonne idée, à condition qu'il soit réaliste de prouver cette fois.

Quelle traduction préférez vous ?

rali

Classement de systèmes

S1	lts-LATL
S2	KIT-Phrase-Based-SMT-System
S3	LIMSI-Ncode-Constrained-Primary
S4	LIUM-EN-FR-12
S5	PROMT-DeepHybrid
S6	RWTH-Jane2-HPBT-constrained
S7	jhu-hiero
S8	onlineA
S9	onlineB
S10	onlineC
S11	onlineD
S12	onlineE
S13	onlineF
S14	uedin-wmt12
S15	uk-dan-moses

Les données viennent de la compétition organisée dans le cadre de WMT2012 (`http://www.statmt.org/wmt12/results.html`)

rali

WMT 2016 [Bojar et al., 2016]

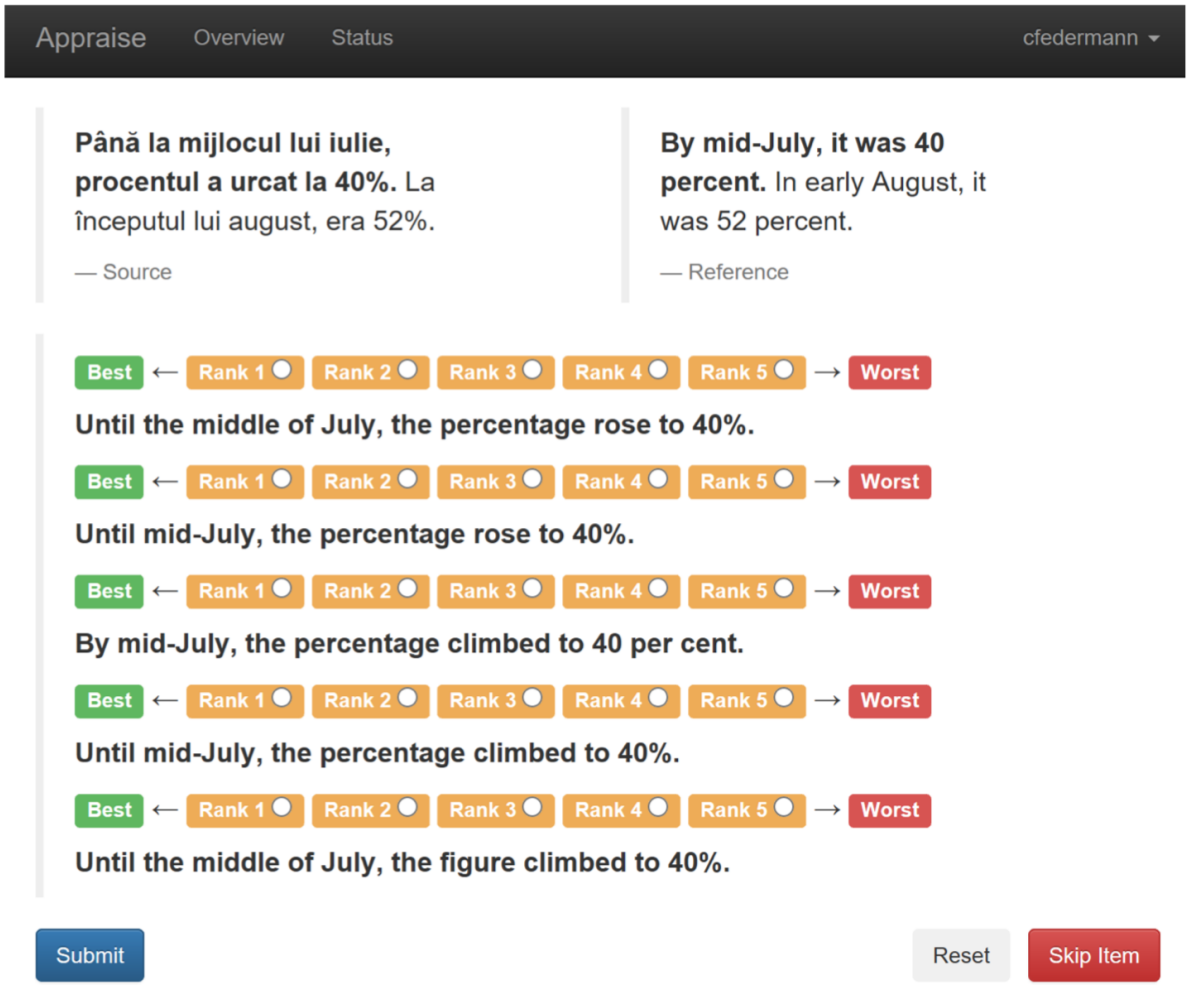


Figure 2: Screenshot of the Appraise interface used in the human evaluation campaign. The annotator is presented with a source segment, a reference translation, and up to five outputs from competing systems (anonymized and displayed in random order), and is asked to rank these according to their translation quality, with ties allowed.

WMT 2016 [Bojar et al., 2016]

Czech–English

#	score	range	system
1	0.62	1	UEDIN-NMT
2	0.32	2	JHU-PBMT
3	0.21	3	ONLINE-B
4	0.11	4-6	TT-BLEU-MIRA
	0.10	4-7	TT-AFRL
	0.09	4-7	TT-NRC-NNBLEU
	0.07	5-8	TT-NRC-MEANT
	0.03	7-10	TT-BEER-PRO
	0.00	8-10	PJATK
	0.00	8-10	TT-BLEU-MERT
5	−0.07	11	ONLINE-A
6	−1.48	12	CU-MRGTREES

English–Czech

#	score	range	system
1	0.59	1	UEDIN-NMT
2	0.43	2	NYU-MONTREAL
3	0.34	3	JHU-PBMT
4	0.30	4-5	CU-CHIMERA
	0.30	4-5	CU-TAMCHYNA
5	0.22	6-7	UEDIN-CU-SYTX
	0.19	6-7	ONLINE-B
6	0.16	8-11	TT-BLEU-MIRA
	0.15	8-12	TT-BEER-PRO
	0.15	8-13	TT-BLEU-MERT
	0.14	9-14	TT-AFRL2
	0.14	9-14	TT-AFRL1
	0.13	9-14	TT-DCU

German–English

#	score	range	system
1	0.82	1	UEDIN-NMT
2	0.25	2-5	ONLINE-B
	0.21	2-5	ONLINE-A
	0.19	2-5	UEDIN-SYNTAX
	0.18	2-6	KIT
	0.04	5-7	UEDIN-PBMT
	0.03	6-7	JHU-PBMT
3	−0.12	8	ONLINE-G
4	−0.67	9	JHU-SYNTAX
5	−0.93	10	ONLINE-F

Russian–English

#	score	range	system
1	0.45	1-2	AMU-UEDIN
	0.43	1-3	ONLINE-G
	0.33	2-4	NRC
	0.25	3-5	ONLINE-B
	0.16	4-5	UEDIN-NMT
2	0.04	6-7	ONLINE-A
	0.02	6-7	AFRL-MITLL-PHR
3	−0.11	8-9	AFRL-MITLL-CNTR
	−0.17	8-9	PROMT-RULE
4	−1.39	10	ONLINE-F

English–Russian

#	score	range	system
1	0.79	1	PROMT-RULE
2	0.30	2-4	AMU-UEDIN

English–German

#	score	range	system
1	0.49	1	UEDIN-NMT
2	0.40	2	METAMIND
3	0.29	3	UEDIN-SYNTAX
4	0.17	4	NYU-MONTREAL
5	−0.01	5-10	ONLINE-B
	−0.01	5-10	KIT-LIMSI
	−0.02	5-10	CAMBRIDGE
	−0.02	5-10	ONLINE-A
	−0.03	5-10	PROMT-RULE
	−0.05	6-10	KIT
6	−0.14	11-12	JHU-SYNTAX
	−0.15	11-12	JHU-PBMT
7	−0.26	13-14	UEDIN-PBMT
	−0.33	13-15	ONLINE-F
	−0.34	14-15	ONLINE-G

Finnish–English

#	score	range	system
1	0.42	1-4	UEDIN-PBMT
	0.40	1-4	ONLINE-G
	0.39	1-4	ONLINE-B
	0.34	1-4	UH-OPUS
2	0.01	5	PROMT-SMT
3	−0.11	6-7	UH-FACTORED
	−0.13	6-7	UEDIN-SYNTAX
4	−0.29	8	ONLINE-A
5	−1.03	9	JHU-PBMT

rali

WMT 2016 [Bojar et al., 2016]

	0.19	6-7	ONLINE-B
6	0.16	8-11	TT-BLEU-MIRA
	0.15	8-12	TT-BEER-PRO
	0.15	8-13	TT-BLEU-MERT
	0.14	9-14	TT-AFRL2
	0.14	9-14	TT-AFRL1
	0.13	9-14	TT-DCU
	0.13	11-14	TT-FJFI
7	0.08	15	ONLINE-A
8	−0.03	16	CU-TECTOMT
9	−0.43	17	TT-USAAR-HMM-MERT
10	−0.54	18	CU-MRGTREES
11	−1.13	19	TT-USAAR-HMM-MIRA
12	−1.33	20	TT-USAAR-HARM

Romanian–English

#	score	range	system
1	0.58	1-2	ONLINE-B
	0.38	1-2	UEDIN-NMT
2	0.10	3	UEDIN-PBMT
3	−0.09	4-5	UEDIN-SYNTAX
	−0.19	4-6	ONLINE-A
	−0.32	5-7	JHU-PBMT
	−0.46	6-7	LIMSI

English–Romanian

#	score	range	system
1	0.45	1-2	UEDIN-NMT
	0.43	1-2	QT21-HIML-COMB
2	0.20	3-7	KIT
	0.16	3-7	UEDIN-PBMT
	0.14	3-7	ONLINE-B
	−0.17	8-9	PROMT-RULE
4	−1.39	10	ONLINE-F

English–Russian

#	score	range	system
1	0.79	1	PROMT-RULE
2	0.30	2-4	AMU-UEDIN
	0.26	2-5	ONLINE-B
	0.26	2-5	UEDIN-NMT
	0.20	3-5	ONLINE-G
3	0.10	6	NYU-MONTREAL
4	−0.02	7-8	JHU-PBMT
	−0.07	7-10	LIMSI
	−0.10	8-10	ONLINE-A
	−0.15	9-10	AFRL-MITLL-PHR
5	−0.31	11	AFRL-MITLL-VERB
6	−1.26	12	ONLINE-F

Turkish–English

#	score	range	system
1	0.82	1-2	ONLINE-B
	0.65	1-3	ONLINE-G
	0.56	2-3	ONLINE-A
2	0.21	4-5	TBTK-SYSCOMB
	0.12	4-6	PROMT-SMT
	−0.00	5-6	YSDA
3	−0.67	7-8	JHU-SYNTAX
	−0.76	7-9	JHU-PBMT
	−0.94	8-9	PARFDA

	0.34	1-4	UH-OPUS
2	0.01	5	PROMT-SMT
3	−0.11	6-7	UH-FACTORED
	−0.13	6-7	UEDIN-SYNTAX
4	−0.29	8	ONLINE-A
5	−1.03	9	JHU-PBMT

English–Finnish

#	score	range	system
1	0.36	1-3	ONLINE-G
	0.31	1-4	ABUMATRAN-NMT
	0.29	1-4	ONLINE-B
	0.23	3-5	ABUMATRAN-CMB
	0.16	4-5	UH-OPUS
2	−0.01	6-8	ABUMATRAN-PB
	−0.02	6-8	NYU-MONTREAL
	−0.02	6-8	ONLINE-A
3	−0.14	9-10	JHU-PBMT
	−0.23	9-12	UH-FACTORED
	−0.28	10-13	AALTO
	−0.30	10-13	JHU-HLTCOE
	−0.35	11-13	UUT

English–Turkish

#	score	range	system
1	0.76	1-2	ONLINE-G
	0.62	1-2	ONLINE-B
2	0.38	3	ONLINE-A
3	0.06	4	YSDA
4	−0.13	5-6	JHU-HLTCOE
	−0.19	5-7	TBTK-MORPH
	−0.29	6-7	CMU

WER (Word Error Rate)

- distance d'édition normalisée (par la taille source / la taille cible / le nombre total d'opérations d'édition)

SRC	She does not think that the issue of British Gas' structure will be a long-term problem .
REF	Elle ne croit pas que le problème structurel de la British perdure .
SMT	Elle ne croit que la question de British Gas' structures sera un problème .

```
REF  : Elle ne  croit pas que le  problème structurel de  la  Brit
CAN  : Elle ne  croit      que la  question                  de
ALI  : <=>  <=> <=>   DEL <=> SUB SUB           DEL          <=> I
```

- ED = 10, 17 opérations $\Rightarrow$ WER = $100 \times 10/17 = 58.8\%$

mWER (Minimum Word Error Rate)

- WER avec la plus proche des références

SRC	She does not think that the issue of British Gas' structure will be a long-term problem .
REF	Elle ne croit pas que le problème structurel de la British perdure .
REF	Elle ne croit pas que la question structurelle de la British Gas sera un problème à long term .
SMT	Elle ne croit que la question de British Gas' structures sera un problème .

```
Elle ne  croit pas que la  question structurelle de  la  British G
Elle ne  croit      que la  question                  de      B
<=>  <=> <=>   DEL <=> <=> <=>              DEL          <=> DEL <=
```

- ED = 8, 20 opérations $\Rightarrow$ mWER = $100 \times 8/20 = 40\%$

PER (Position Independant Rate)

- toute permutation d'une référence a un taux d'erreur null.

SRC	She does not think that the issue of British Gas' structure will be a long-term problem .
REF	Elle ne croit pas que le problème structurel de la British perdure .
SMT1	Elle ne croit que la question de British Gas' structures sera un problème .
SMT2	Gas' structures Elle ne croit sera un que la question de British problème .

- PER = $100 \times 8/14 = 57.14\%$
- SER (Sentence Error Rate) = $100 \times \frac{I - \sum_i^I \delta(Trad_i, Ref_i)}{I} = 100\%$

BLEU

[Papineni et al., 2002]

Ref	Can you give me some medicine for headache ?
SMT1	Can I have some medicine for headache ?
SMT2	Can you give me prescribe some medicine ?
SMT3	Can I have some medicine for headache ?

très bon

bon

ok

nul

BLEU

- Comparaison de n-grams entre une traduction et une ou plusieurs références
- 1g jusqu'à 4g (typiquement)
- $BLEU = BP \times \exp(\sum_{n=1}^{N} w_n \frac{|n\text{-grams}_{trad} \cap n\text{-grams}_{ref}|}{|n\text{-grams}_{trad}|})$

  moyenne pondérée des précisions n-gram (entre 0 et 1, 1 = parfait)
- BP (Brevity Penalty) = $\min(1, \exp(|trad|/|ref|))$

  pénaliser les phrases trop courtes par rapport à la référence
- $w_n = 1/N$

BLEU

- valable au niveau d'un texte (500 phrases ou plus)
- la présence d'un n-gram (n grand) est très bien payée
- semble corréler avec les jugements humains (note globale entre 1 et 5)
- les n-grams les plus fréquents sont souvent les moins informatifs (fluency plus qu'adequacy)
- lire [Doddington, 2002] pour une variante biaisée davantage vers la fidélité (adequacy) : NIST

rali

Les premiers pas. . .

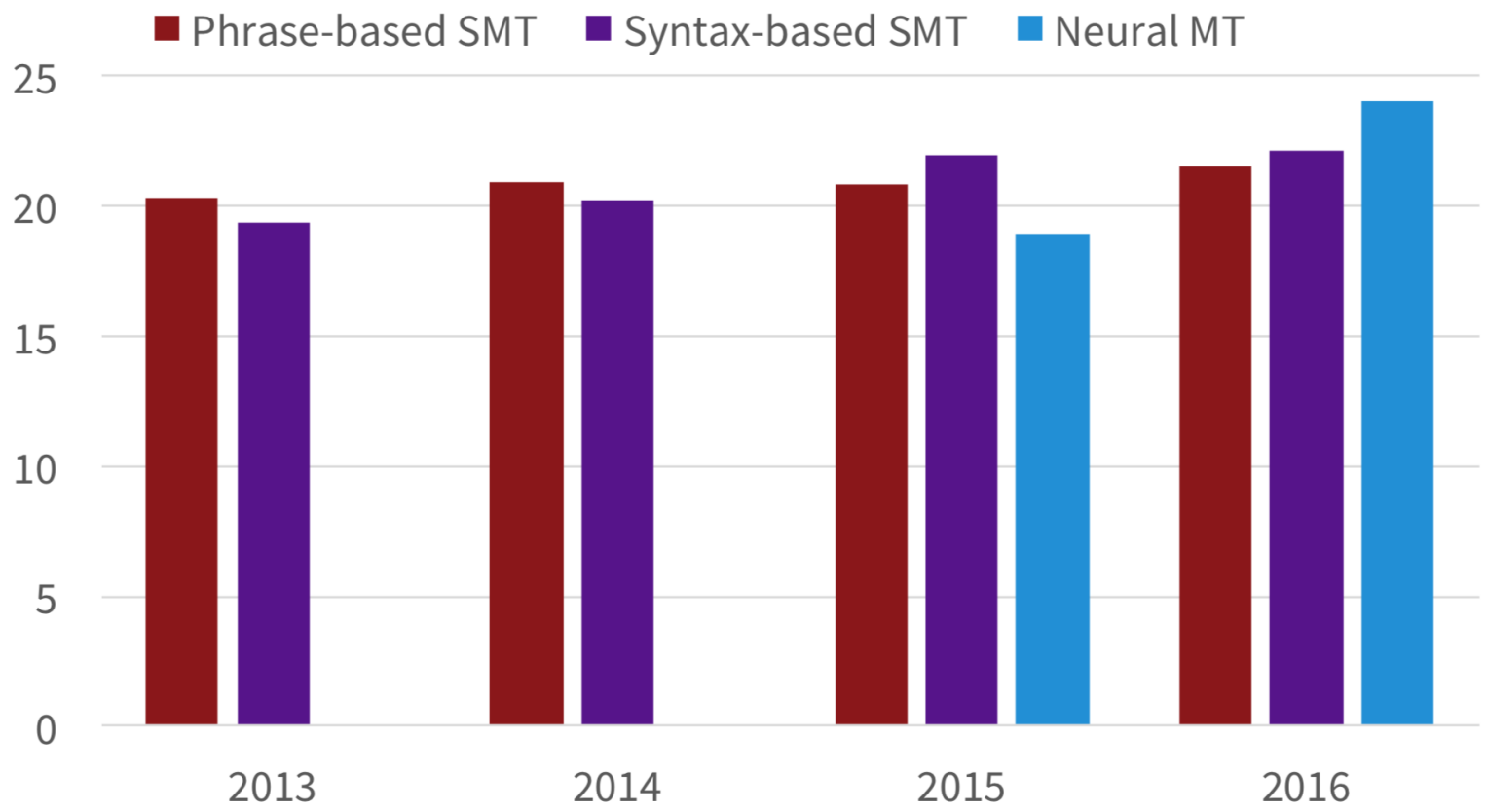
- Modèle de langue neuronal “à la Bengio” via rescoring *[Schwenk et al., 2006, Le et al., 2012b, Le et al., 2012a]*
- Modèle neuronal global $p(\text{trg word}|\text{src sentence})$ embarqué à même un système SMT *[Patry and Langlais, 2011]*
- Le pionnier des systèmes encodeur-décodeur (Allen :1987) 31/40 mots (EN/SP)

rali

https://sites.google.com/site/acl16nmt/home

Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



From [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

uedin-nmt	34.2
metamind	32.3
NYU-UMontreal	30.8
cambridge	30.6
uedin-syntax	30.6
KIT/LIMSI	29.1
KIT	29.0
uedin-pbmt	28.4
jhu-syntax	26.6

EN→DE

uedin-nmt	38.6
uedin-pbmt	35.1
jhu-pbmt	34.5
uedin-syntax	34.4
KIT	33.9
jhu-syntax	31.0

DE→EN

uedin-nmt	25.8
NYU-UMontreal	23.6
jhu-pbmt	23.6
cu-chimera	21.0
uedin-cu-syntax	20.9
cu-tamchyna	20.8
cu-TectoMT	14.7
cu-mergedtrees	8.2

EN→CS

uedin-nmt	31.4
jhu-pbmt	30.4
PJATK	28.3
cu-mergedtrees	13.3

CS→EN

uedin-pbmt	35.2
uedin-nmt	33.9
uedin-syntax	33.6
jhu-pbmt	32.2
LIMSI	31.0

RO→EN

QT21-HimL-SysComb	28.9
uedin-nmt	28.1
RWTH-SYSCOMB	27.1
uedin-pbmt	26.8
uedin-lmu-hiero	25.9
KIT	25.8
lmu-cuni	24.3
LIMSI	23.9
jhu-pbmt	23.5
usfd-rescoring	23.1

EN→RO

uedin-nmt	26.0
amu-uedin	25.3
jhu-pbmt	24.0
LIMSI	23.6
AFRL-MITLL	23.5
NYU-UMontreal	23.1
AFRL-MITLL-verb-annot	20.9

EN→RU

amu-uedin	29.1
NRC	29.1
uedin-nmt	28.0
AFRL-MITLL	27.6
AFRL-MITLL-contrast	27.0

RU→EN

- Edinburgh NMT

[Patry and Langlais, 2011]

- tirage de Bernouilli de chaque mot du vocabulaire

$$\Pr(\mathbf{t} \mid \mathbf{s}) = \overbrace{\prod_{t \in \mathbf{t}} \Pr(t \mid \mathbf{s})}^{\text{Present}} \overbrace{\prod_{t \in \mathcal{T} - \mathbf{t}} 1 - \Pr(t \mid \mathbf{s})}^{\text{Absent}} \quad (1)$$

  - $\mathcal{T}$ (resp. $\mathcal{S}$) vocabulaire cible (resp. source)
- $\Pr(t \mid \mathbf{s})$ estimé par un perceptron multicouche (MLP) :

$$\vec{h} = \tanh(W\vec{s}) \quad (2)$$

$$\vec{y} = \text{tsigmoid}(V\vec{h}) \quad (3)$$

$$\text{tsigmoid}(z) = \text{sigmoid}(z - 4.6) \quad (4)$$

  - $W \in \mathbb{R}^{d \times |\mathcal{S}|}$ et $V \in \mathbb{R}^{|\mathcal{T}| \times d}$ les paramètres du modèle
  - $\vec{h}$ la couche cachée ($\in \mathbb{R}^d$), $\vec{s}$ un vecteur one-hot ($\in \mathbb{R}^{|\mathcal{S}|}$)
  - $\vec{y}$ le vecteur de probabilités ($\in \mathbb{R}^{|\mathcal{T}|}$)

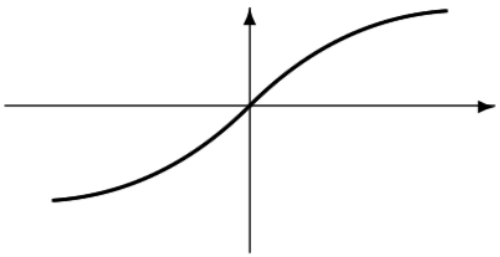
rali

Fonctions d'activation classiques

Hyperbolic tangent

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

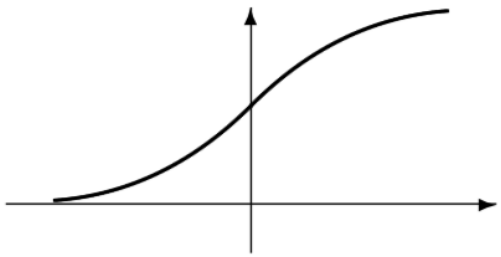
output ranges
from –1 to +1



Logistic function

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$$

output ranges
from 0 to +1



Rectified linear unit

$$\text{relu}(x) = \max(0,x)$$

output ranges
from 0 to ∞

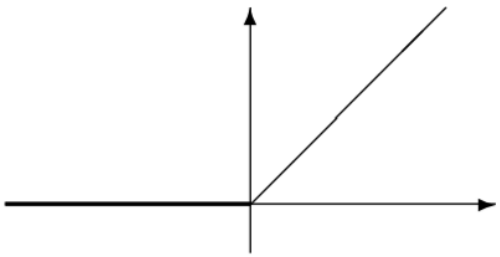


Figure 13.3: Typical activation functions in neural networks.

https ://arxiv.org/abs/1709.07809

[Patry and Langlais, 2011] : Motivation

Source	Target
the floor of the plant	le plancher de l'usine
the stem of the plant	la tige de la plante
the leaves and stem of the flower	les feuilles et la tige de la fleur
the floors and walls of the house	les planchers et les murs de la maison

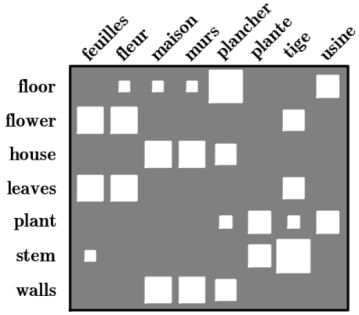


Figure 3: Weights learned for a logistic regression model optimized by gradient descent.

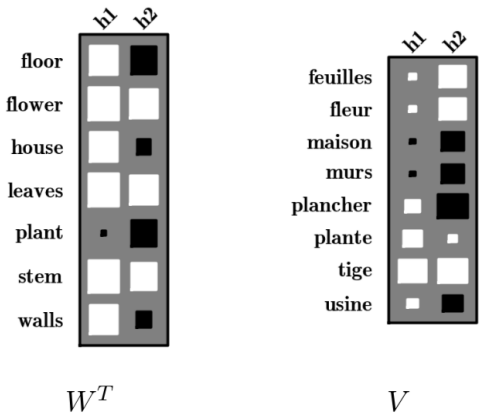


Figure 4: Weights learned for a MLP.

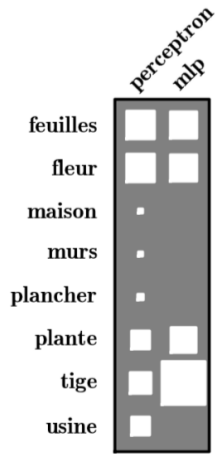


Figure 5: Predictions of the logistic regression model and the MLP for *the leaves of this plant*.

- carré blanc/noir : poids positif/négatif
- taille proportionnelle au poids

[Patry and Langlais, 2011] : Entraînement

- entraîné par descente de gradient (custom) de manière à maximiser la log vraisemblance d'un corpus d'entraînement

- mini-batches de 100 paires de phrases
- line-search sur la fonction d'erreur régularisée pour ajuster le learning rate (0.5, 0.05, 0.01)
- 20 epoch

[Patry and Langlais, 2011] : Entraînement

- mots pleins (noms, verbes, etc.) apparaissant au moins 20 fois dans le TRAIN :
  - $\mathcal{S}$ (FR) : 143 980 $\rightarrow$ 28 095
  - $\mathcal{T}$ (EN) : 133 141 $\rightarrow$ 21 915
- comparaison de 3 modèles de prédiction :

IBM1	IBM1 entraîné avec GIZA++ [Och and Ney, 2003].
PERCEP	équivalent à de la regression logistique (pas de couche cachée)
MLP-64	MLP avec 64 unités cachées

[Patry and Langlais, 2011] : Intégration

- 2 scores ajoutés à un pipeline similaire à Moses (in house) :
  - forme simplifiée de l'équation 1 *[Mauser et al., 2009]*

$$odd(\mathbf{t}|\mathbf{s}) = \prod_{t \in \mathbf{t}} \frac{y_{t|\mathbf{s}}}{1 - y_{t|\mathbf{s}}} \tag{5}$$

    - un odd de 5 indique par exemple qu'un mot a 5 fois plus de chance d'être présent dans la traduction que d'en être absent
  - nombres de mots avec une "bonne" prédiction :

$$pred(\mathbf{t} \mid \mathbf{s}) = \left|\{t \in \mathbf{t} \mid y_{t|\mathbf{s}} > \alpha\}\right| \tag{6}$$

- scores faciles à intégrer dans Moses (incrémentaux)

[Patry and Langlais, 2011] : Évaluation

In-domain

System	BLEU (%) *odd*	BLEU (%) *pred*
baseline	30.06	30.06
+ IBM1	30.32	30.65
+ PERCEP	30.68	30.71
+ MLP-64	**30.86**	**31.00**

Out-domain

System	BLEU (%) *odd*	BLEU (%) *pred*
baseline	19.05	19.05
+ IBM1	20.00	**19.92**
+ PERCEP	19.93	**19.82**
+ MLP-64	**20.45**	**19.89**

- FR-EN, WMT 2009 (http://www.statmt.org/wmt09/)
  - TRAIN : EUROPARL + NEWSCOMMENTARY
  - DEV : TEST2007 TEST : NEWSTEST2009

Modèle récurrent

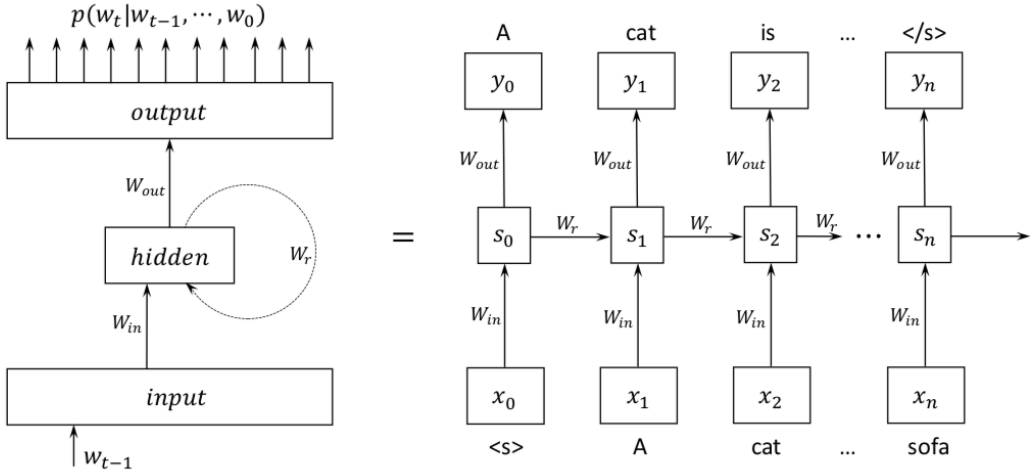


Figure 1: The network architecture of a standard RNNLM and its unrolled version for an example input sentence: <s> A cat is sitting on a sofa </s>.

fig. prise de : [Ji et al., 2015]

- $s_t = \sigma(W_i^T \times x_t + W_r \times s_{t-1} + b_i) \in \mathbb{R}^h$
  $W_i \in \mathbb{R}^{|V| \times h}$, $W_i \in \mathbb{R}^{h \times h}$, $b_i \in \mathbb{R}^h$, $\sigma(x) = 1/(1 + \exp(-x))$
- $y_t = f(W_o \times s_t + b_o) \in \mathbb{R}^{|V|}$
  $W_o \in \mathbb{R}^{|V| \times h}$, $b_o \in \mathbb{R}^{|V|}$, $f(v_i) = \exp(v_i)/\sum_j \exp(v_j)$

Cas d'un modèle de langue

- x_t est un one-hot-vector ($\in \mathbb{R}^{|V|}$ avec des 0 partout sauf pour la dimension représentant le mot qui est à 1)

chat	$(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, \ldots)^T$
chien	$(0, 0, 0, 0, 0, 0, 1, 0, 0, 0, \ldots)^T$

- $\theta = (W_i, W_o, W_r, b_i, b_o)$ sont les paramètres du modèle

[Mikolov et al., 2010]

- a **largement** contribué à la popularité des RNN en traitement des langues

	PPL
KN5	93.7
feedforward NN	85.1
recurrent NN	80.0
4$\times$RNN + KN5	73.5

Switchboard (4M de mots) — conversations téléphoniques

[Mikolov et al., 2010]

	PPL		WER	
Model	RNN	RNN+KN	RNN	RNN+KN
KN5 - baseline	-	221	-	13.5
RNN 60/20	229	186	13.2	12.6
RNN 90/10	202	173	12.8	12.2
RNN 250/5	173	155	12.3	11.7
RNN 250/2	176	156	12.0	11.9
RNN 400/10	171	152	12.5	12.1
3xRNN static	151	143	11.6	11.3
3xRNN dynamic	128	121	11.3	11.1

- RNN configuration is written as *hidden/threshold* - 90/10 means that network has 90 neurons in hidden layer and threshold for keeping words in vocabulary is 10.

- 37M de mots du NYT (6.4M seulement pour les RNNs)
  - plusieurs semaines d'entraînement
- RNN + KN = combinaison linéaire (0.75 sur le RNN)

BiLSTM

- les réseaux récurrents sont difficiles à entraîner *[Bengio et al., 1994]*
- solutions `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`
  - LSTM *[Hochreiter and Schmidhuber, 1997]*
  - GRU *[Chung et al., 2014]*
- Si le contexte droit est pertinent pour la prédiction, on peut également avoir une couche qui va de la droite vers la gauche
  - baptisé BiLSTM *[Graves et al., 2013]*
  - terriblement populaire

rali

BiLSTM

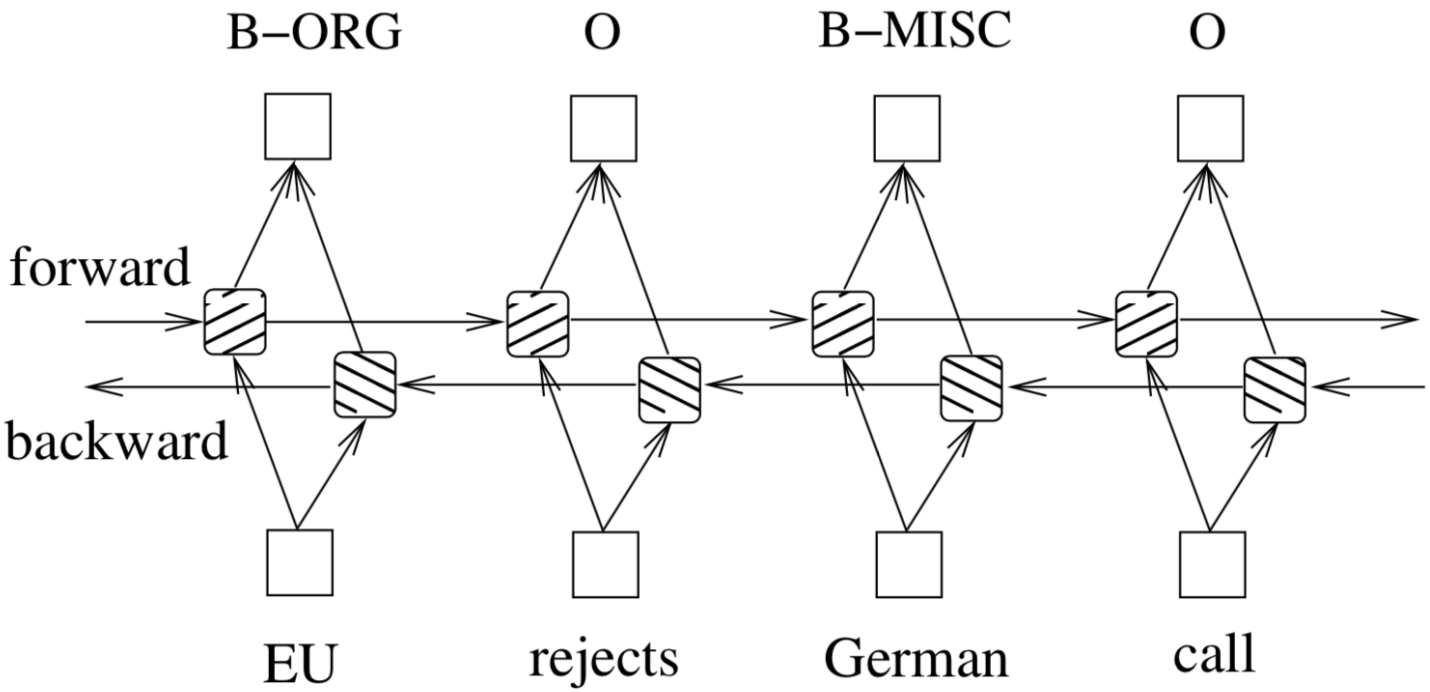
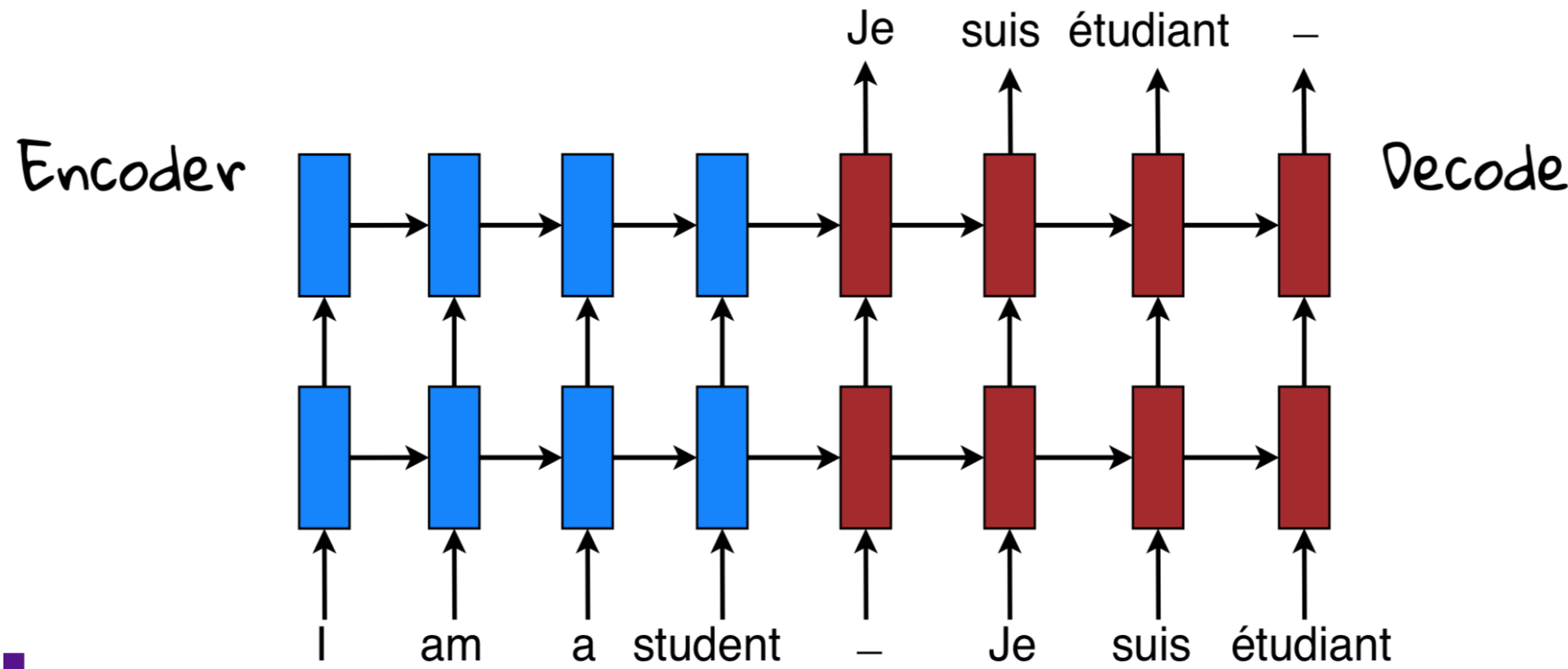


Figure 4: A bidirectional LSTM network.

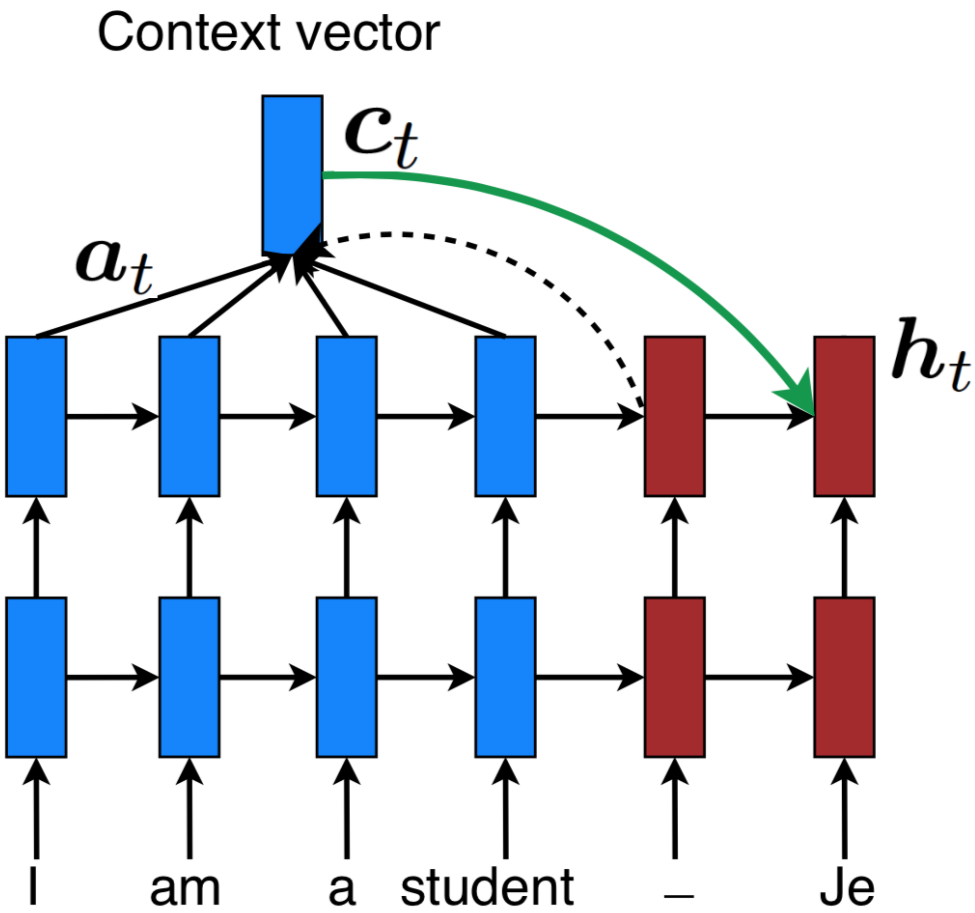
fig. prise de [Huang et al., 2015]

Traduction neuronale : un modèle de langue !



https://sites.google.com/site/acl16nmt/home

Mécanisme d'attention *[Bahdanau et al., 2014]*



https://sites.google.com/site/acl16nmt/home

Simple ?

- systèmes longs à entraîner (normalisation softmax)
- modèles peu efficaces à prédire les mots rares (vocabulaire souvent restreint pour des raisons calculatoires)
- beaucoup de solutions à ces problèmes (et d'autres) dans les 3 dernières années
- l'une d'elle est remarquable de simplicité : BPE
  [Sennrich et al., 2015]
- lire l'excellent tutoriel de Luong, Cho et Manning (ACL 2016)

rali

Byte Pair Encoding *[Sennrich et al., 2015]*

- A word segmentation algorithm:
  - Start with a vocabulary of characters.
  - Most frequent ngram pairs ↦ a new ngram.

Dictionary

5 **lo** w
2 **lo** w e r
6 n e w est
3 w i d est

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, est, **lo**

Add a pair (l, o) with freq 7

166

(Example from Sennrich)

https://sites.google.com/site/acl16nmt/home

- Meilleur système à WMT 2016 (BPE + BiLSTM)

[Isabelle et al., 2017]

- Création de jeux de tests à problème, un problème par test (afin que l'annotation soit simple à conduire)

Source	The repeated calls from his mother should have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
System	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct? (y/n) **Yes**	

  - 108 tests (phrases)
  - annotation manuelle des sorties de traduction statistique, neuronale, google MT et DeepL
- Très convainquant quant aux avantages des systèmes neuronaux
- La suite dans les acétates des auteurs

[Koehn and Knowles, 2017]

domain adaptation

System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 32.8	45.1 42.2	35.3 44.7	17.9 17.9	26.4 20.8
Law	31.1 34.4	12.1 18.2	3.5 6.9	1.3 2.2	2.8 6.0
Medical	3.9 10.2	39.4 43.5	2.0 8.5	0.6 2.0	1.4 5.8
IT	1.9 3.7	6.5 5.3	42.1 39.8	1.8 1.6	3.9 4.7
Koran	0.4 1.8	0.0 2.1	0.0 2.3	15.9 18.8	1.0 5.5
Subtitles	7.0 9.9	9.3 17.8	9.2 13.6	9.0 8.4	25.9 22.1

Figure 1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, NMT systems (left bars) show more degraded performance out of domain.

rali

[Koehn and Knowles, 2017]

taille du TRAIN

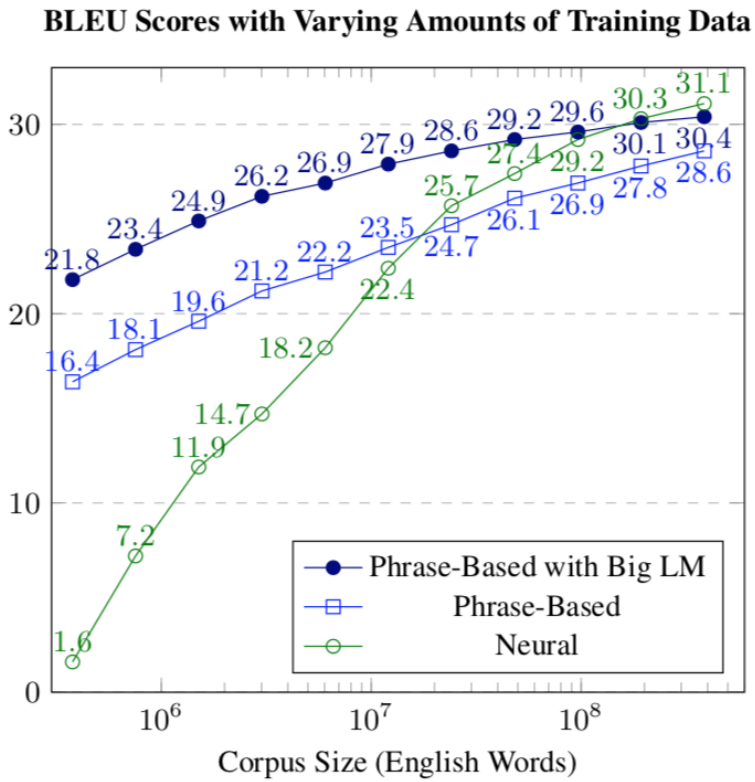


Figure 3: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data. Quality for NMT starts

[Koehn and Knowles, 2017]

phrases longues

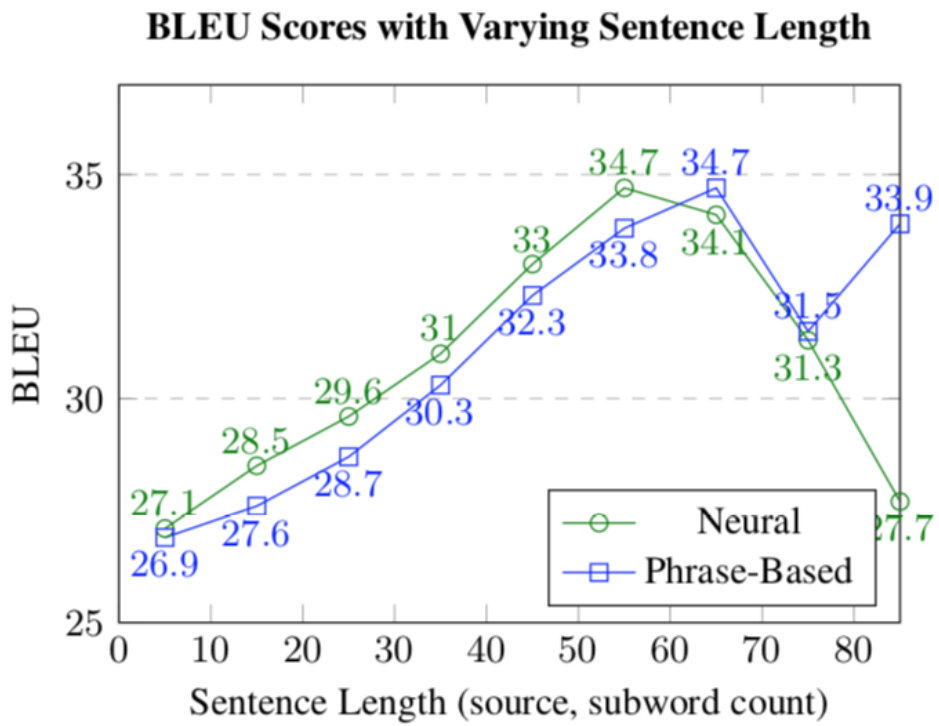


Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sen-

[Koehn and Knowles, 2017]

décodage

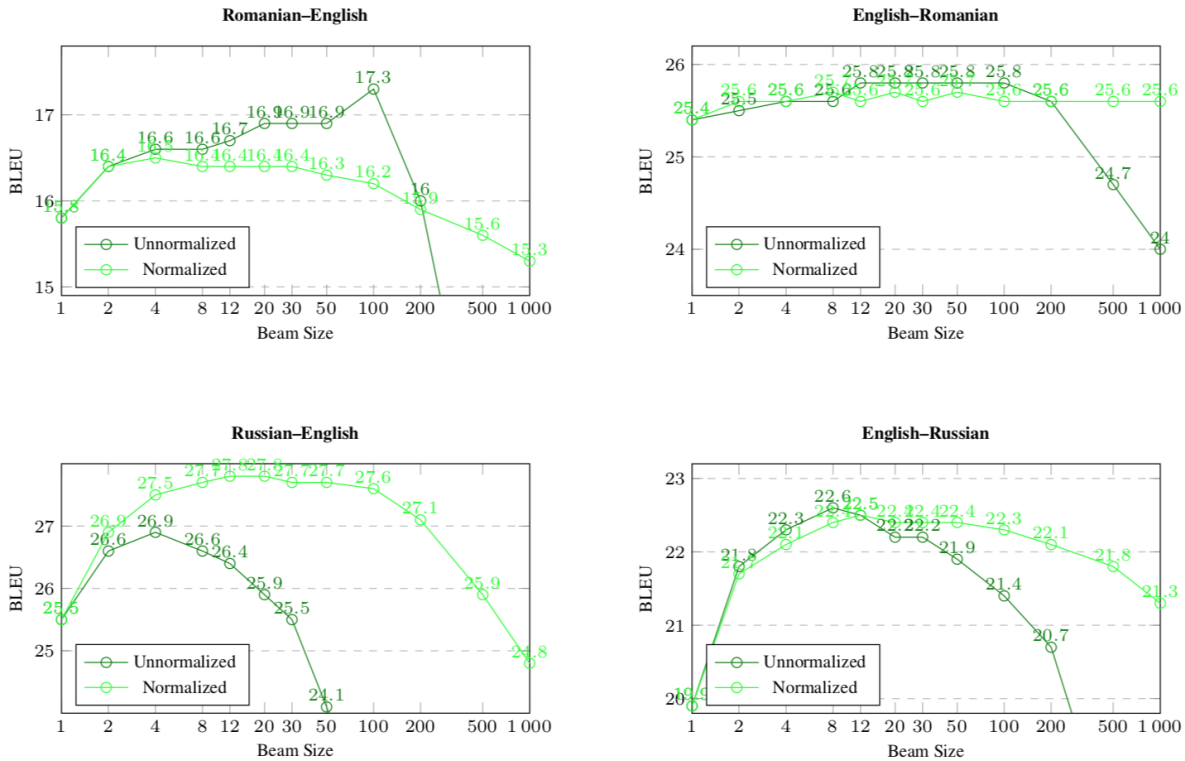


Figure 10: Translation quality with varying beam sizes. For large beams, quality decreases, especially when not normalizing scores by sentence length.

- problème : des traductions trop courtes sont produites

ACL-35 (1997).
Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), Madrid, Spain.

Bahdanau, D., Cho, K., and Bengio, Y. (2014).
Neural machine translation by jointly learning to align and translate.
CoRR, abs/1409.0473.

Bengio, Y., Simard, P., and Frasconi, P. (1994).
Learning long-term dependencies with gradient descent is difficult.
Trans. Neur. Netw., 5(2) :157–166.

Bisson, F. and Fluhr, C. (2000).
Sentence alignment in bilingual corpora based on crosslingual querying.
In *Computer-Assisted Information Retrieval*, RIAO 2000, pages 529–542.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016).
Findings of the 2016 conference on machine translation.
In *Proceedings of the First Conference on Machine Translation*, pages 131–198.

Braune, F. and Fraser, A. (2010).
Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora.
In *23rd COLING*, pages 81–89.

Brown, P., Lai, J., and Mercer, R. (1991).
Aligning Sentences in Parallel Corpora.
In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 169–176, Berkeley, California.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993).
The mathematics of statistical machine translation : Parameter estimation.
Computational Linguistics, 19(2) :263–311.

Cer, D., Jurafsky, D., and Manning, C. D. (2008).
Regularization and search for minimum error rate training.
In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 26–34.

Chang, J. and Chen, M. (1997).
An alignment method for noisy parallel corpora based on image processing techniques.
In [ACL-35, 1997], pages 297–304.

Cherry, C. and Foster, G. (2012).
Batch tuning strategies for statistical machine translation.
In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics :*

Human Language Technologies, NAACL HLT '12, pages 427–436.

Chiang, D. (2007).
Hierarchical phrase-based translation.
Comput. Linguist., 33(2) :201–228.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014).
Empirical evaluation of gated recurrent neural networks on sequence modeling.
CoRR, abs/1412.3555.

Collins, M. (1999).
Head-Driven Statistical Models for Natural Language Parsing.
PhD thesis, University of Pennsylvania.

Debili, F. (1992).
Aligning sentences in bilingual texts french-english and french-arabic.
In *Proceedings of the International Conference on Computational Linguistics (COLING) 1992*, pages 517–525, Nantes, France.

rali

Deng, Y., Kumar, S., and Byrne, W. (2007).
Segmentation and alignment of parallel text for statistical machine translation.
Natural Language Engineering, 13 :235–260.

Doddington, G. (2002).
Automatic evaluation of machine transslation quality using n-gram cooccurrence statistics.
In *Proceedings of the HLT*, pages 257–258, San Diego, USA.

Fung, P. and Church, K. (1994).
K-vec : A new approach for aligning sentences in bilingual corpora.
In *Proceedings of the International Conference on Computational Linguistics (COLING) 1994*, pages 1096–1102, Kyoto, Japan.

Gale, W. A. and Church, K. W. (1993a).
A program for aligning sentences in bilingual corpora.
In *Computational Linguistics*, volume 19, pages 75–102.

Gale, W. A. and Church, K. W. (1993b).
A program for aligning sentences in bilingual corpora.
Comput. Linguist., 19(1) :75–102.

Goutte, C., Carpuat, M., and Foster, G. (2012).
The impact of sentence alignment errors on phrase-based machine translation performance.
In *10th AMTA*.

Graves, A., Mohamed, A., and Hinton, G. E. (2013).
Speech recognition with deep recurrent neural networks.
CoRR, abs/1303.5778.

Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.
Neural computation, 9 :1735–1780.

Huang, Z., Xu, W., and Yu, K. (2015).
Bidirectional LSTM-CRF models for sequence tagging.
CoRR, abs/1508.01991.

Hutchins, W. J. and Somers, H. L. (1992).

An Introduction to Machine Translation.
Academic Press.

Isabelle, P., Cherry, C., and Foster, G. (2017).
A challenge set approach to evaluating machine translation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Ji, S., Vishwanathan, S. V. N., Satish, N., Anderson, M. J., and Dubey, P. (2015).
Blackout : Speeding up recurrent neural network language models with very large vocabularies.
CoRR, abs/1511.06909.

Jurafsky, D. and Martin, J. H. (2000).
Speech and Language Processing.
Prentice Hall.

Kay, M. and Röscheisen, M. (1993).
Text-translation alignment.
Computational Linguistics, 19(1) :121–142.

rali

Knight, K. (1999a).
Decoding complexity in word-replacement translation models.
Computational Linguistics, 25(4).

Knight, K. (1999b).
A statistical mt tutorial workbook.
http ://www.isi.edu/natural-language/people/knight.html.

Koehn, P. and Hoang, H. (2007).
Factored translation models.
In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

Koehn, P. and Knowles, R. (2017).
Six challenges for neural machine translation.
CoRR, abs/1706.03872.

Koehn, P., Och, F., and Marcu, D. (2003).
Statistical phrase-based translation.

In *Proceedings of the Human Language Technology Conference (HLT)*, pages 127–133.

Lamraoui, F. and Langlais, P. (2013).
Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment ?
In *XIV Machine Translation Summit*, pages 77–84, Nice, France.

Langé, J.-M. and Gaussier, E. (1995).
Alignement de corpus multilingues au niveau des phrases.
T.A.L, 36(1-2) :67–80.

Langlais, P., Simard, M., and Véronis, J. (1998).
Methods and practical issues in evaluating alignment techniques.
In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, Montréal, Canada.

Le, H.-S., Allauzen, A., and Yvon, F. (2012a).
Continuous space translation models with neural networks.
In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics :*

Human Language Technologies, pages 39–48, Montréal, Canada. Association for Computational Linguistics.

Le, H.-S., Allauzen, A., and Yvon, F. (2012b).
Measuring the influence of long range dependencies with neural network language models.
In *Proceedings of the NAACL-HLT 2012 Workshop : Will We Ever Really Replace the N-gram Model ? On the Future of Language Modeling for HLT*, pages 1–10, Montréal, Canada.

Li, P., Sun, M., and Xue, P. (2010).
Fast-champollion : a fast and robust sentence alignment algorithm.
In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 710–718.

Manning, C. D. and Schütze, H. (1999).
Foundations of Statistical Natural Language Processing.
MIT Press.

Mauser, A., Hasan, S., and Ney, H. (2009).

rali

Extending statistical machine translation with discriminative and trigger-based lexicon models.
In *Conference on Empirical Methods in Natural Language Processing*.

Melamed, I. D. (1997).
A portable algorithm for mapping bitext correspondence.
In [ACL-35, 1997], pages xx–yyy.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010).
Recurrent neural network based language model.
In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, volume 2, pages 1045–1048.

Moore, R. C. (2002).
Fast and accurate sentence alignment of bilingual corpora.
In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation : From Research to Real Users*, AMTA '02, pages 135–144.

Och, F. J. (2003).
Minimum error rate training in statistical machine translation.
In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167.

Och, F. J. and Ney, H. (2003).
A systematic comparison of various statistical alignment models.

Computational Linguistics, 29(1) :19–51.

Och, F. J. and Ney, H. (2004).
The alignment template approach to statistical machine translation.
Comput. Linguist., 30(4) :417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).
Bleu : a method for automatic evaluation of machine translation.
In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

rali

Patry, A. and Langlais, P. (2011).
Going beyond word cooccurrences in global lexical selection for statistical machine translation using a multilayer perceptron.
In *5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 658–666, Chiang Mai, Thailand.

Quan, X., Kit, C., and Song, Y. (2013).
Non-monotonic sentence alignment via semisupervised learning.
In *ACL*, pages 622–630.

Ribeiro, A., Dias, G., Lopes, G., and Mexia, J. (2001).
Cognates alignment.
In *Machine Translation Summit VIII, Machine Translation in The Information Age*, pages 287–292, Santiago de Compostela, Galicia, Spain.

Schwenk, H., Dechelotte, D., and Gauvain, J.-L. (2006).
Continuous space language models for statistical machine translation.

In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 723–730, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015).
Neural machine translation of rare words with subword units.
CoRR, abs/1508.07909.

Sennrich, R. and Volk, M. (2011).
Iterative, mt-based sentence alignment of parallel texts.
In *18th Nordic Conference of Computational Linguistics*.

Simard, M., Foster, G., and Isabelle, P. (1992).
Using cognates to align sentences in bilingual corpora.
In *Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 67–81, Montréal, Québec.

Simard, M. and Plamondon, P. (1996).
Bilingual sentence alignment : Balancing robustness and accuracy.

rali

In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*, Montréal, Québec.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005).
Parallel corpora for medium density languages.
In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Véronis, J. and Langlais, P. (2000).
Evaluation of parallel text alignment systems : The ARCADE project, volume 13, chapter 19, pages 369–388.
Parallel Text Processing, Kluwer.

Vogel, S., Ney, H., and Tillmann, C. (1996).
Hmm-based word alignement in statistical translation.
In *Proceedings of the International Conference on Computational Linguistics (COLING) 1996*, pages 836–841, Copenhagen, Denmark.

Wu, D. (1994).

rali

Aligning a parallel english-chinese corpus statistically with lexical criteria.
In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87, Las Cruces, New Mexico.

Wu, D. (1997).
Stochastic inversion transduction grammars and bilingual parsing of parallel corpora.
Computational Linguistics, 23(3) :377–404.

Wu, D. (2000).
Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars, volume 13, chapter 19, pages 139–168.
Parallel Text Processing, Kluwer.

Yamada, K. and Knight, K. (2001).
A syntax-based statistical translation model.

In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 523–530, Toulouse, France.

Yu, Q., Max, A., and Yvon, F. (2012).
Revisiting sentence alignment algorithms for alignment visualization and evaluation.
In *5th workshop BUCC*, pages 10–16.