Alignement de phrases

Pierre Zweigenbaum

LIMSI, CNRS, Université Paris-Saclay
pz@limsi.fr — https://perso.limsi.fr/pz/

Plan

Le problème de l'alignement de phrases

Méthodes

- Auxiliaire : segmentation en phrases
- Principe : similarité de structure des textes
- Principe : similarité des phrases
- Méthodes pour appliquer les principes

Outils disponibles

- Char_align (Gale & Church)
- GMA (Melamed)
- En deux passes
- Alignement par apprentissage supervisé

Le problème de l'alignement de phrases

- Le cas idéal : alignement 1-1

anglais	français
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.

d'après (Gale & Church, 1993)

Le parallélisme n'est pas toujours strict : 2-1

- Alignement 2-1

anglais	français
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Le parallélisme n'est pas toujours strict : 2-2

- Alignement 2-2

anglais	français
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.

Autres cas de non-parallélisme des phrases

- (Alignement 1-1)
- Alignement 2-1
- Alignement 2-2
- Alignement . . .

Attitudes vis-à-vis du non-parallélisme

Des objectifs différents

- Collecte de traductions de mots et d'expressions
  - par exemple pour la traduction automatique
  - rechercher les alignements les plus fiables (1-1)
- Construction d'un bitexte complet
  - par exemple pour la lecture de livres bilingues
  - chercher à aligner toutes les phrases source et cible

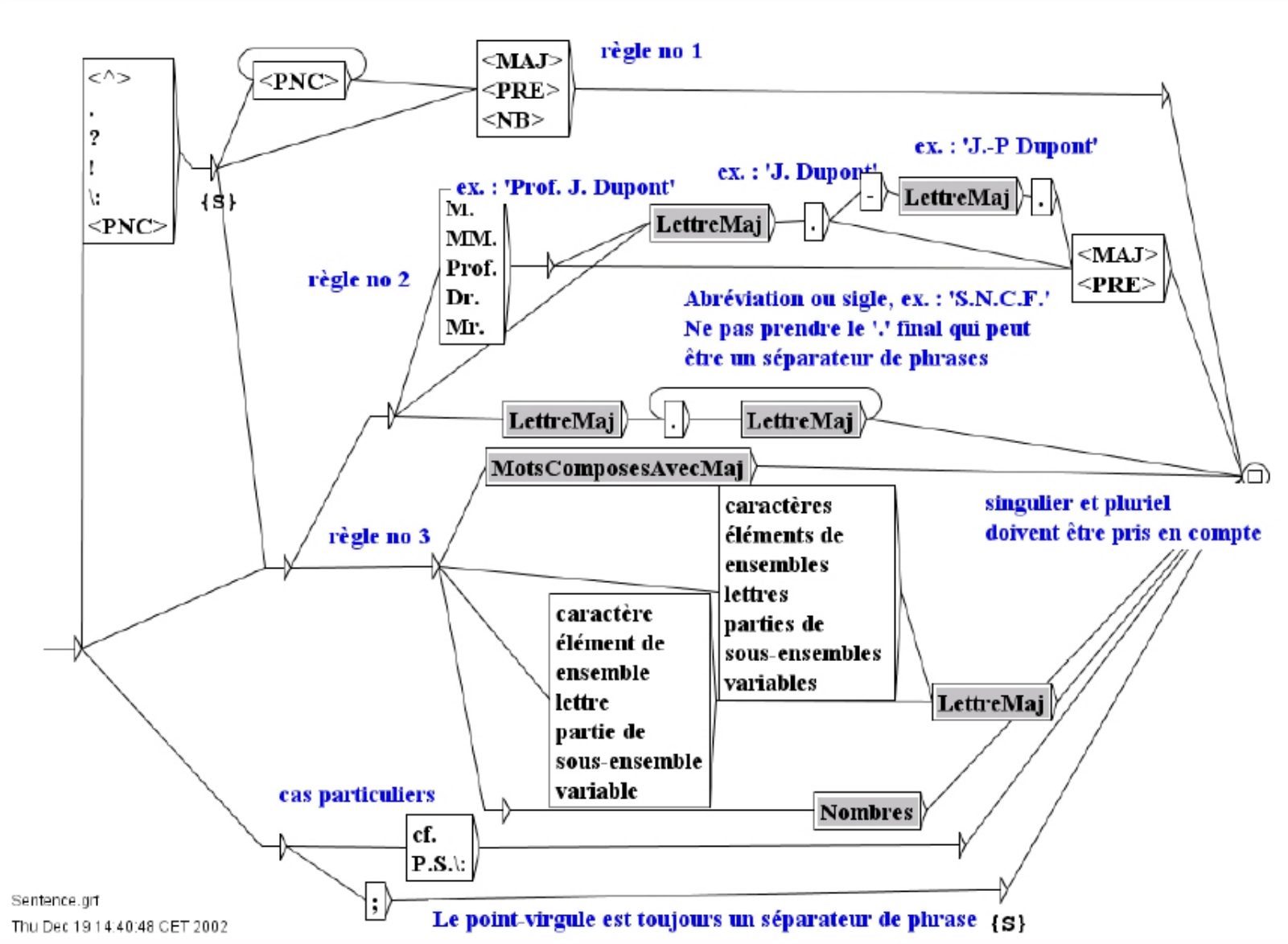
Auxiliaire : segmentation en phrases

- Règles
  - Unitex
  - Europarl tools
  - Perl Lingua::Sentence
  - LingPipe
- Apprentissage supervisé
  - NLTK
  - Stanford CoreNLP

Règles de segmentation

- Segmentation sur les ponctuations de fin de phrase
  . ! ?
- Problèmes :
  - Ambiguïté du point
    abréviation, point décimal en anglais
  - Autres ponctuations ?
    : ; « ([ ]) »
  - La disposition du texte peut segmenter :
    paragraphes, titres, alinéas, tableaux
- Voir par exemple
  - `http://www.statmt.org/europarl/v7/tools.tgz`
  - Module Perl Lingua::Sentence
    `http://search.cpan.org/~achimru/Lingua-Sentence-1.05/lib/Lingua/Sentence.pm`

Automate d'Unitex pour la segmentation en phrases



Segmentation en phrases supervisée

Apprentissage supervisé des frontières de phrase

- Tâche, au choix :
  - catégoriser chaque espace : frontière de phrase ou pas
  - catégoriser chaque mot : dernier mot d'une phrase ou pas
- Corpus d'entraînement
  - Texte(s) où les frontières de phrase sont marquées
- Corpus de test
  - Texte brut

Principes d'alignement de phrases

Comment savoir quelles phrases se correspondent ?

- Similarité de structure des textes
  - structure hiérarchique (paragraphes...)
  - régularité de l'ordre des phrases
- Similarité des phrases
  - forme
  - contenu lexical

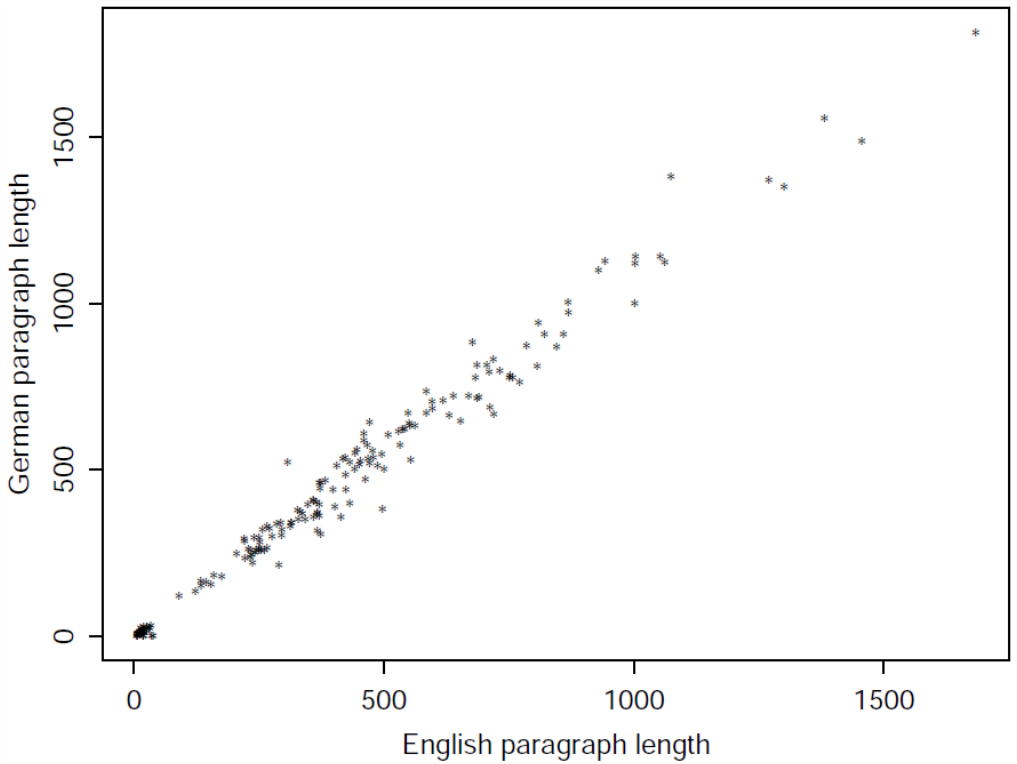
Similarité de structure hiérarchique des textes

Aligner d'abord les paragraphes

- Voire même les différentes divisions d'un texte :
- Parties, chapitres, sections, etc.
  - cf alignement de documents (X)HTML

Observation : longueur des paragraphes

- Les longueurs des paragraphes alignés des deux langues sont dans un rapport approximativement constant

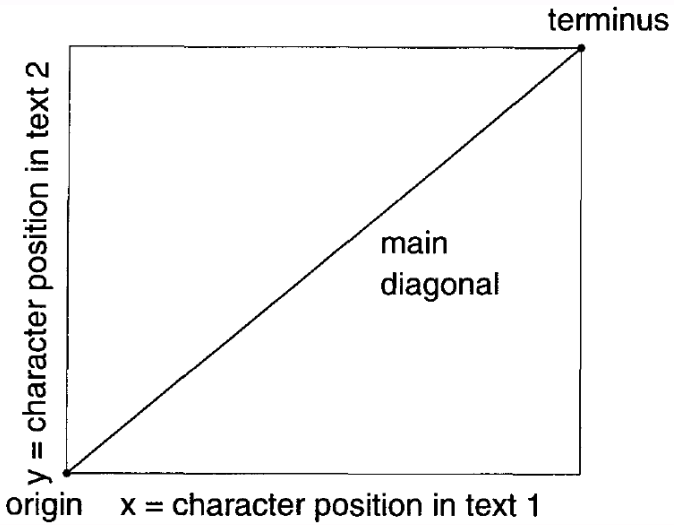


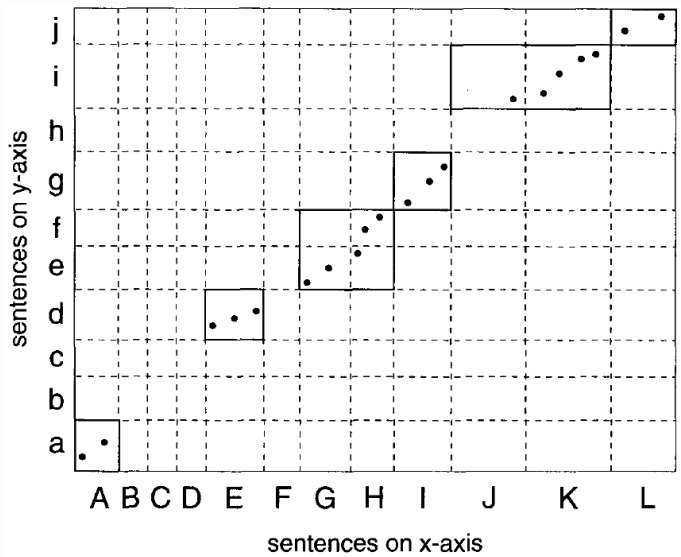
- Un alignement préalable des paragraphes aide (Gale & Church, 1993) [si les documents s'y prêtent]

Régularité de l'ordre des phrases

- Les phrases sont généralement présentées dans le même ordre dans les documents source et cible
- Algorithme de programmation dynamique

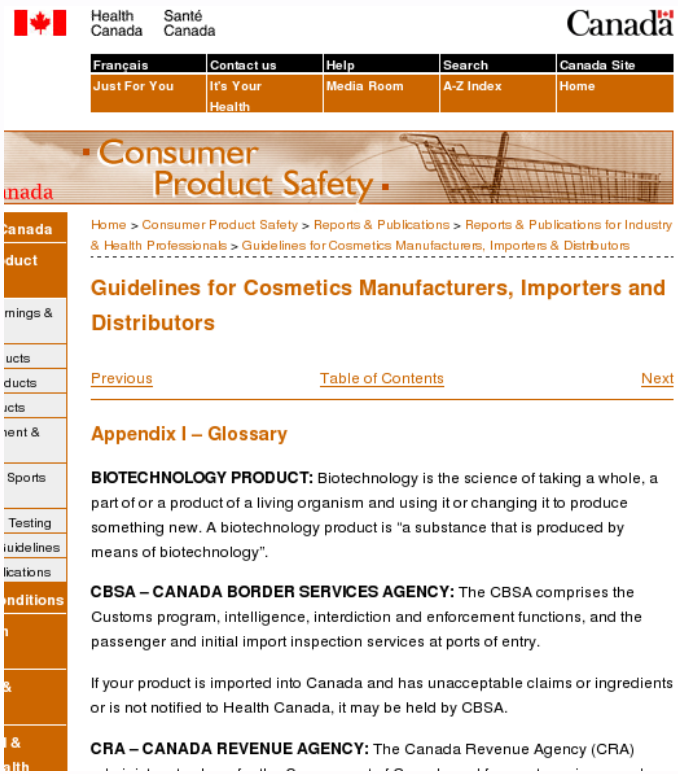
Correspondances entre positions des phrases : couloir autour de la diagonale





d'après *(Melamed, 1999)*

Exception : documents non parallèles (ex : glossaire)


Health Canada Santé Canada

Canada

Français | Contact us | Help | Search | Canada Site
Just For You | It's Your Health | Media Room | A-Z Index | Home

Consumer Product Safety

Home > Consumer Product Safety > Reports & Publications > Reports & Publications for Industry & Health Professionals > Guidelines for Cosmetics Manufacturers, Importers & Distributors

Guidelines for Cosmetics Manufacturers, Importers and Distributors

Previous Table of Contents Next

Appendix I – Glossary

BIOTECHNOLOGY PRODUCT: Biotechnology is the science of taking a whole, a part of or a product of a living organism and using it or changing it to produce something new. A biotechnology product is "a substance that is produced by means of biotechnology".

CBSA – CANADA BORDER SERVICES AGENCY: The CBSA comprises the Customs program, intelligence, interdiction and enforcement functions, and the passenger and initial import inspection services at ports of entry.

If your product is imported into Canada and has unacceptable claims or ingredients or is not notified to Health Canada, it may be held by CBSA.

CRA – CANADA REVENUE AGENCY: The Canada Revenue Agency (CRA)


Santé Canada Health Canada

Canada

English | Contactez-nous | Aide | Recherche | Site du Canada
Spécialement pour vous | Votre santé et vous | Salle des médias | Index A-Z | Accueil

Sécurité des produits de consommation

Accueil > Sécurité des produits de consommation > Rapports et Publications > Rapports et publications pour l'industrie et des professionnels de la santé > Lignes directrices à l'intention des fabricants, importateurs et distributeurs de cosmétiques

Lignes directrices à l'intention des fabricants, importateurs et distributeurs de cosmétiques

Précédente Table des matières Prochaine

Annexe I – Glossaire

ACIA – AGENCE CANADIENNE D'INSPECTION DES ALIMENTS : L'ACIA dispense des programmes d'inspection des aliments, des plantes et des animaux partout au Canada. Son rôle consiste à faire respecter les normes établies par Santé Canada en ce qui concerne la salubrité et la qualité nutritive des aliments, à établir des normes en matière de santé des animaux et de protection des végétaux, à veiller à leur application et à leur respect, puis à assurer les services d'inspection et d'application des règlements.

ALIMENT : « Comprend notamment tout article fabriqué, vendu ou présenté comme pouvant servir de nourriture ou de boisson à l'être humain. Inclut la gomme à mâcher ainsi que tout ingrédient pouvant être mélangé avec un aliment à quelque fin que ce soit. »

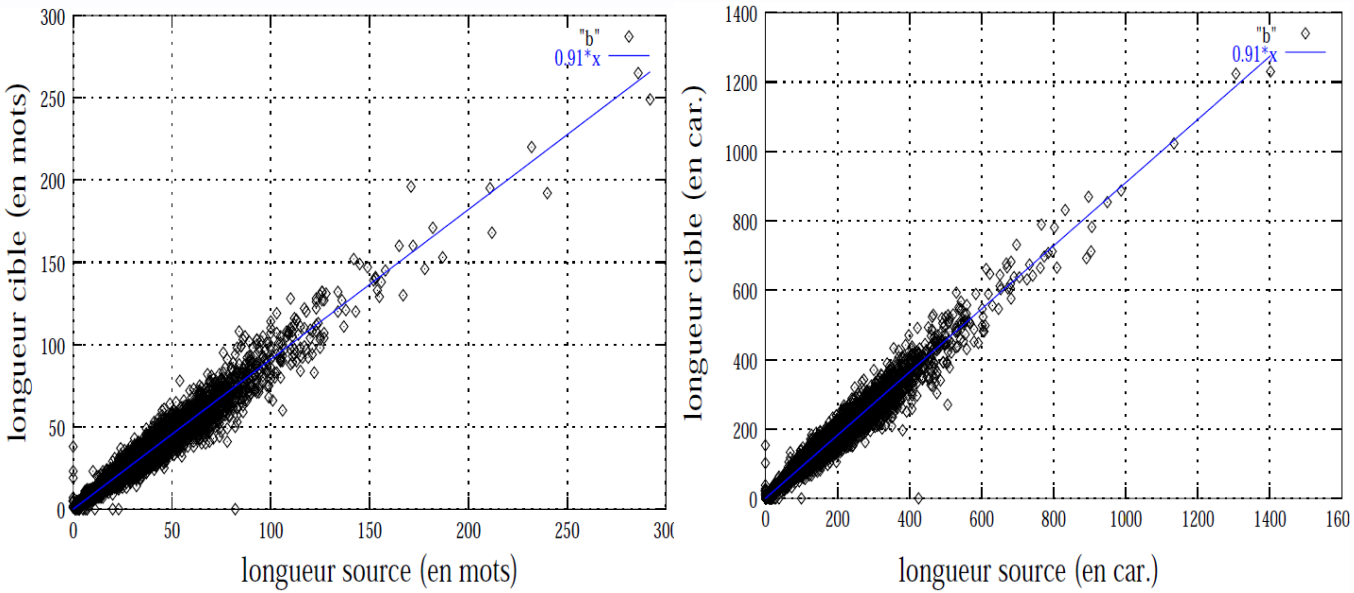
Évaluation ARCADE 1 : le corpus « technique » contenait un glossaire, ce qui a causé de très mauvais résultats d'alignement pour tous les systèmes participants (Véronis & Langlais, 1999)

Similarité des phrases

- Similarité de longueur (en caractères ; en mots)
- Mots communs :
  - directement (nombres, noms propres, ponctuations)
  - approximativement (cognats ; à travers un lexique bilingue)

Observation : longueur des phrases

Des régions de texte plus longues ont tendance à avoir des traductions plus longues



(d'après Langlais, 2005)

- Le test sur la longueur en caractères fonctionne mieux que celui sur la longueur en mots (Gale & Church, 1993)

Mots communs

Certains « mots » se retrouvent tels quels à travers la traduction

- Nombres : partie numérique des dates, quantités monétaires
- Noms propres : souvent identiques
- Signes de ponctuation : parenthèses

Dépend du couple de langues, et repose sur l'identité des écritures

Mots similaires

Cognats : mots de forme proche (« vrais amis »)

- Identité : *table / table*
- « Préfixe » commun suffisamment long (4 lettres) : *activité / activity*
  - attention aux faux amis : *librairie / library*
- Ressemblance : distance d'édition : *gouvernement / government*
- Translittération

Mots traduits

Est-ce que la phrase cible a des chances d'être la traduction de la phrase source ?

- Selon un lexique bilingue fourni
- Selon un lexique de transfert appris sur le corpus (souvent probabilisé)

Méthodes d'alignement

- Application heuristique des principes
- Principes comme composantes d'un score à maximiser
- Principes comme caractéristiques pour un apprentissage supervisé

Outils disponibles

- `Char_align` (Gale & Church, 1993) : longueur des phrases
- GMA/GSA (Melamed, 1999) : mixte, avec cognats, lexique
- (Moore, 2002) : mixte, sans lexique externe

Char_align (Gale & Church, 1993)

- Alignement des *paragraphes*
  - Suppose un alignement 1:1
  - Élimine d'abord les « pseudo-paragraphes » : titre, signature
    - Critère : pseudo-paragraphes généralement < 50 caractères, vrais paragraphes généralement > 100 caractères
- Alignement des *phrases*
  - Critère de similarité : *rapport des longueurs des phrases* (nombre de caractères)
- Pas d'utilisation d'informations lexicales

Le code du programme est dans l'article (Gale & Church, 1993) :
`http://www.aclweb.org/anthology/J93-1004.pdf` (en C), voir aussi
`http://www.statmt.org/europarl/v7/tools.tgz` (`sentence-align-corpus.perl`)

Char_align (Gale & Church, 1993)

- Alignements élémentaires :

alignement	1:1	1:0	0:1	2:1	1:2	2:2
probabilité	0,89	0,0099	0,0099	0,089	0,089	0,011

- Calcule un coût probabiliste pour chaque alignement élémentaire d'une paire de (groupes de) phrases
  - Probabilité a priori de ce type d'alignement
  - Probabilité du rapport des longueurs des deux (groupes de) phrases si alignement
- Détermine l'ensemble d'alignements qui minimise le coût total
  - « Programmation dynamique »
  - Essaie chaque type d'alignement élémentaire à chaque position
    - Conserve le moins coûteux
  - Exploration incrémentale du meilleur alignement global
    - pour un nombre croissant de phrases des deux textes en partant du début de chaque texte

Programmation dynamique : initialisation

Illustration avec le calcul de la distance de Levenshtein entre deux chaînes de caractères

		c	*h*	*i*	*e*	*n*	*s*
	0	1	2	3	4	5	6
n	1						
i	2						
c	3						
h	4						
e	5						

Programmation dynamique : progression (1)

Illustration

		c	*h*	*i*	*e*	*n*	*s*
	0	1	2	3	4	5	6
n	1	1	2	3	4	4	5
i	2						
c	3						
h	4						
e	5						

Programmation dynamique : progression (2)

Illustration

		c	*h*	*i*	*e*	*n*	*s*
	0	1	2	3	4	5	6
n	1	1	2	3	4	4	5
i	2	2	2	2	3	4	5
c	3						
h	4						
e	5						

Programmation dynamique : fin

Illustration

		c	*h*	*i*	*e*	*n*	*s*
	0	1	2	3	4	5	6
n	1	1	2	3	4	4	5
i	2	2	2	2	3	4	5
c	3	2	3	3	3	4	5
h	4	3	2	3	4	4	5
e	5	4	3	3	3	4	5

Chemin de coût minimal

Illustration

		c	*h*	*i*	*e*	*n*	*s*
	0	1	2	3	4	5	6
n	1	1	2	3	4	4	5
i	2	2	2	2	3	4	5
c	3	2	3	3	3	4	5
h	4	3	2	3	4	4	5
e	5	4	3	3	3	4	5

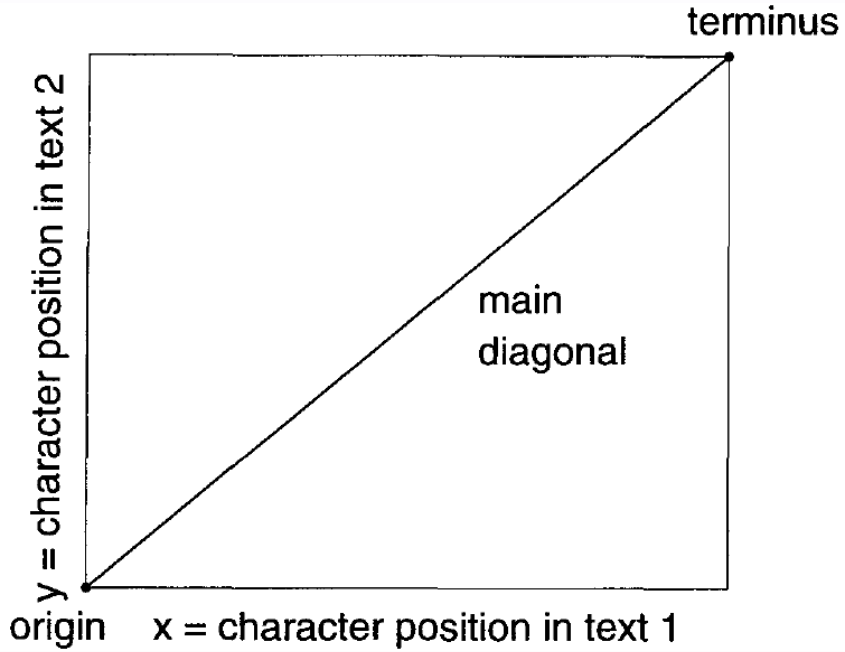
GMA (Melamed, 1999)

(Melamed, 1999)

- Exploite la correspondance entre longueurs des traductions selon une approche géométrique
- Utilise des connaissances lexicales (paramétrables) : cognats, lexique bilingue
- Les contraintes géométriques réduisent le nombre de correspondances lexicales à examiner
- Deux étapes :
  - Identification et sélection de « points de correspondance » (SIMR)
  - Identification de correspondances entre « segments » (GSA : alignement des phrases)

`http://nlp.cs.nyu.edu/GMA/`

Espace du bitexte, axes



- Points de correspondance véritables (« TPC ») :
  - mots, paragraphes, chapitres, alinéas, etc.
- Correspondance de bitexte (« bitext map ») :
  - ensemble de points de correspondance
- Recherche d'une correspondance la plus proche possible de la « véritable »

Correspondances entre mots

Mots identiques (le plus simple)

Cognats orthographiques

Selon le rapport entre la longueur de la plus longue sous-séquence commune (non nécessairement contiguë) et la longueur du mot le plus long

gouvernement / *government* : 10/12
conseil / *conservative* : 6/12

Cognats phonétiques : pour des écritures différentes

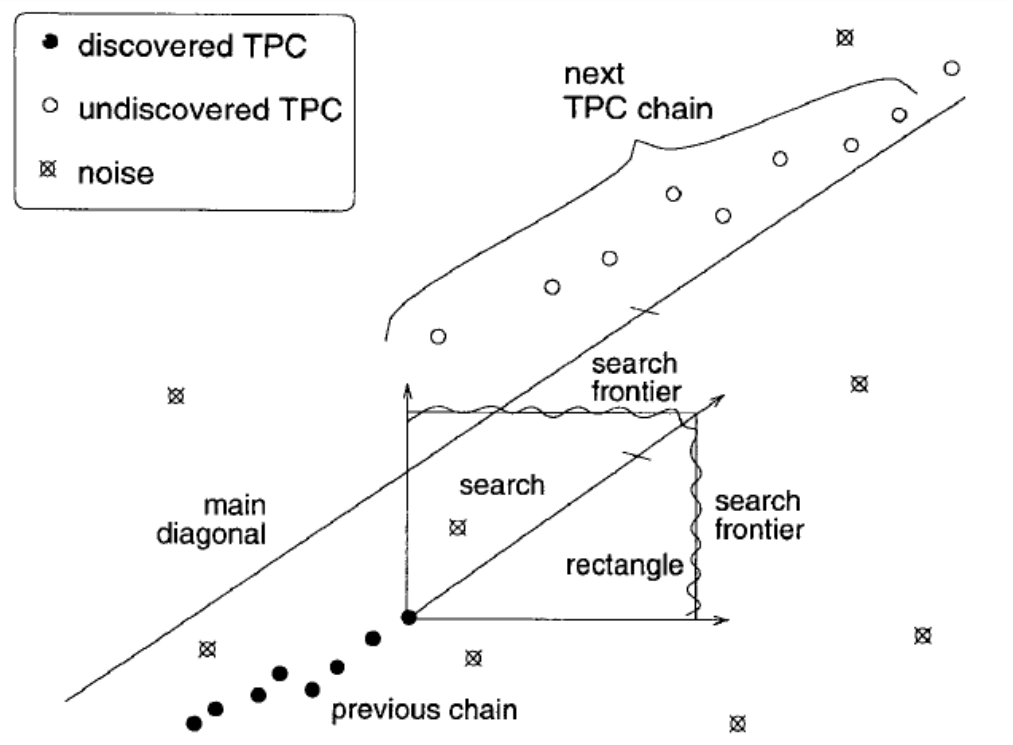
Pourraient être intégrés de la même façon (transducteur)

Correspondances entre mots

- Lexique bilingue d'amorçage
- Listes de mots vides
  - mots grammaticaux : *a*, *an*, *on*, *par*
  - paires de faux-amis : *(librairie*, *library)*

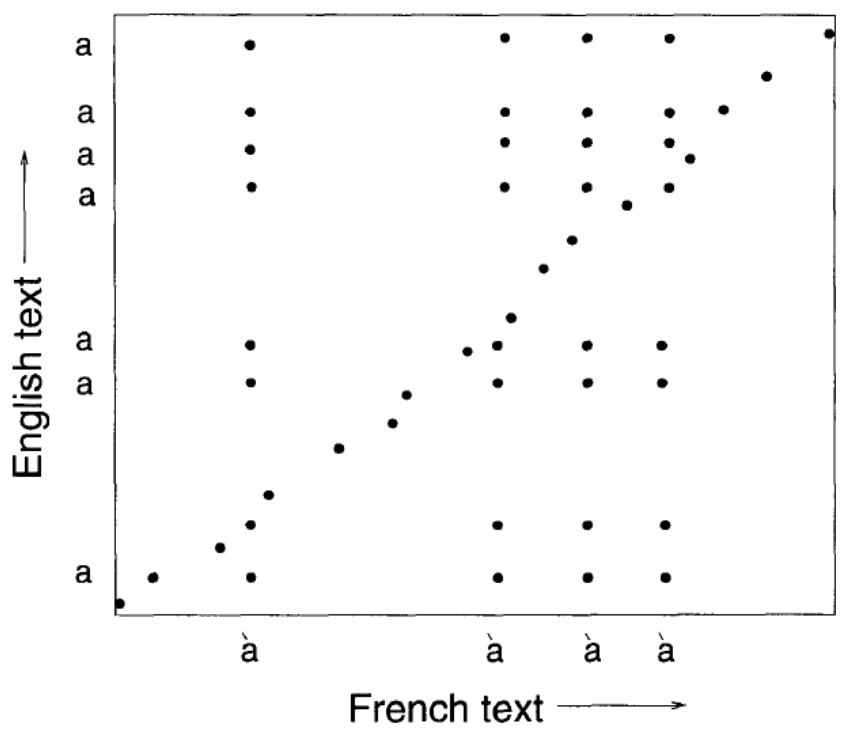
Contraintes sur l'espace de recherche

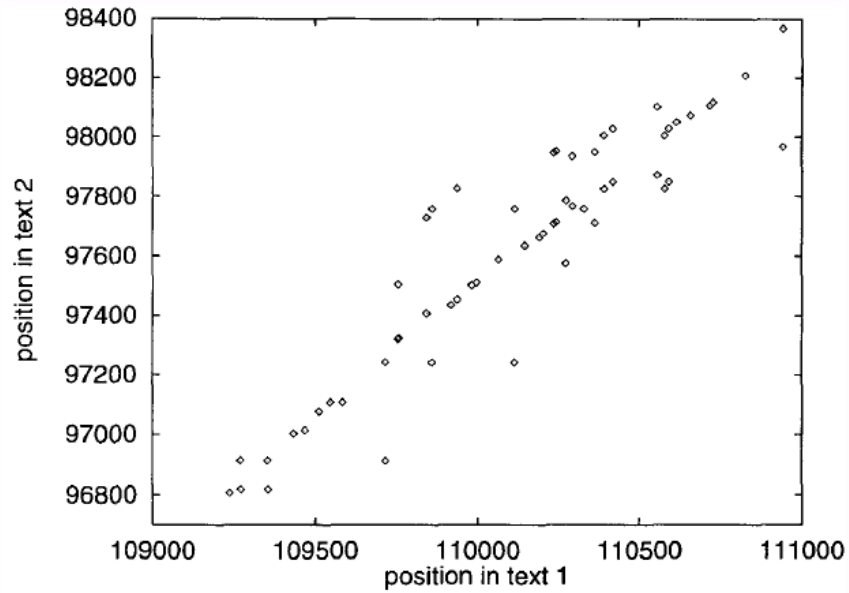
Smooth Injective Map Recognizer : rectangle de recherche



Élimination de correspondances erronées

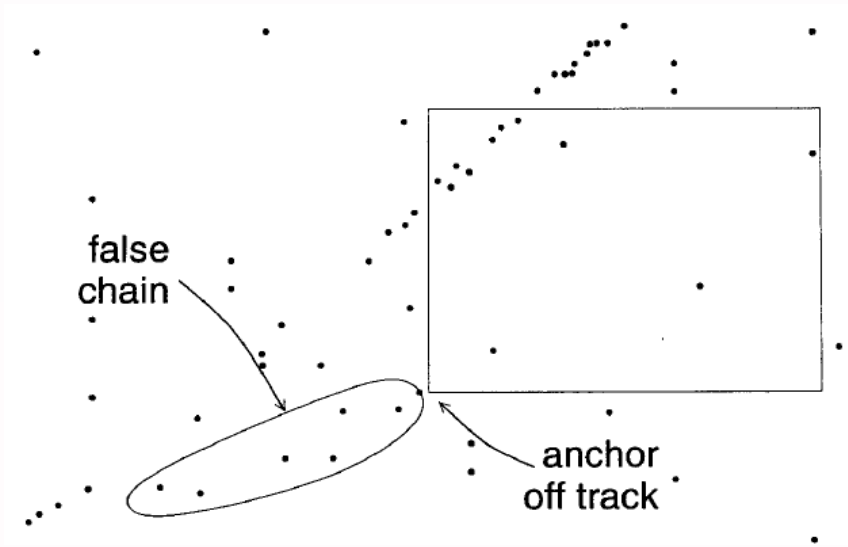
Les mots fréquents provoquent la détection de points de correspondance erronés, qui s'alignent en rangées et en colonnes

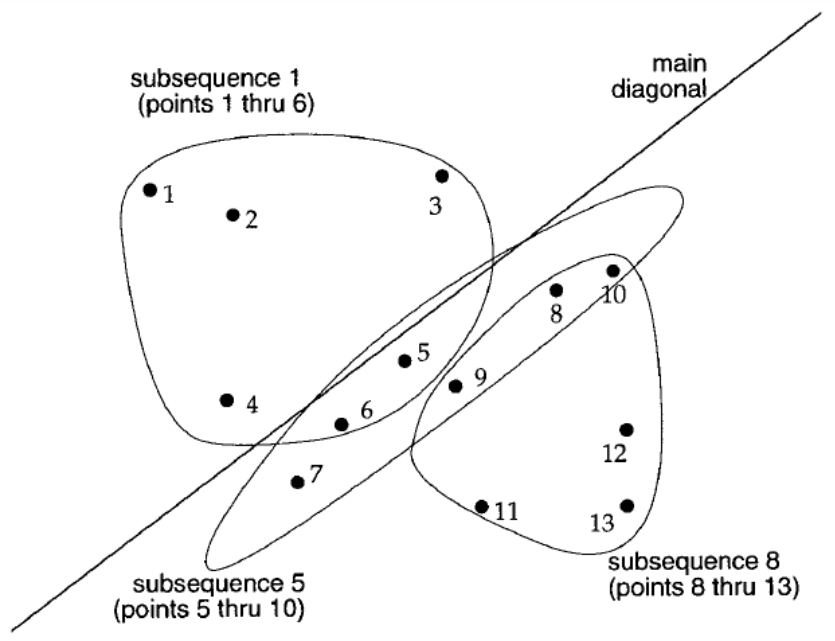




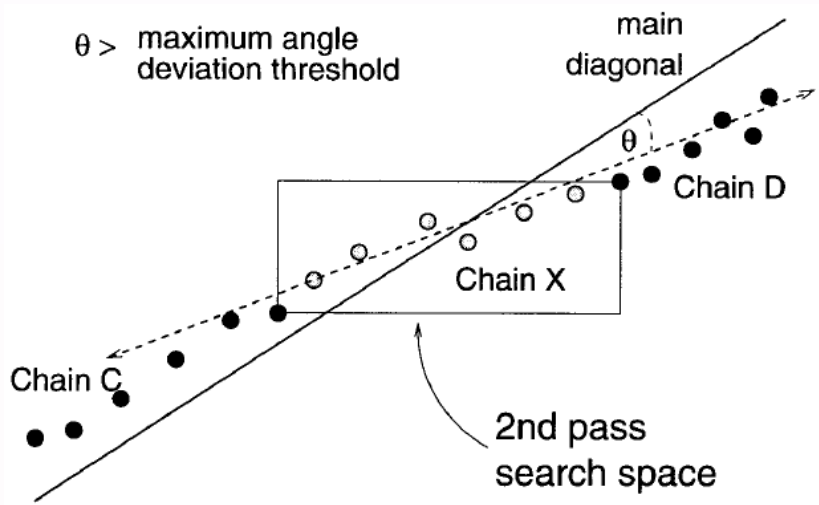
Contraintes sur la sélection des points

- Injectivité : deux points dans la même chaîne de TPC ne peuvent pas avoir le même x ou le même y
- Linéarité : les TPC tendent à s'aligner (ils forment des « chaînes »)
- Faible variance de la pente : la pente d'une chaîne de TPC est rarement très différente de la pente du bitexte

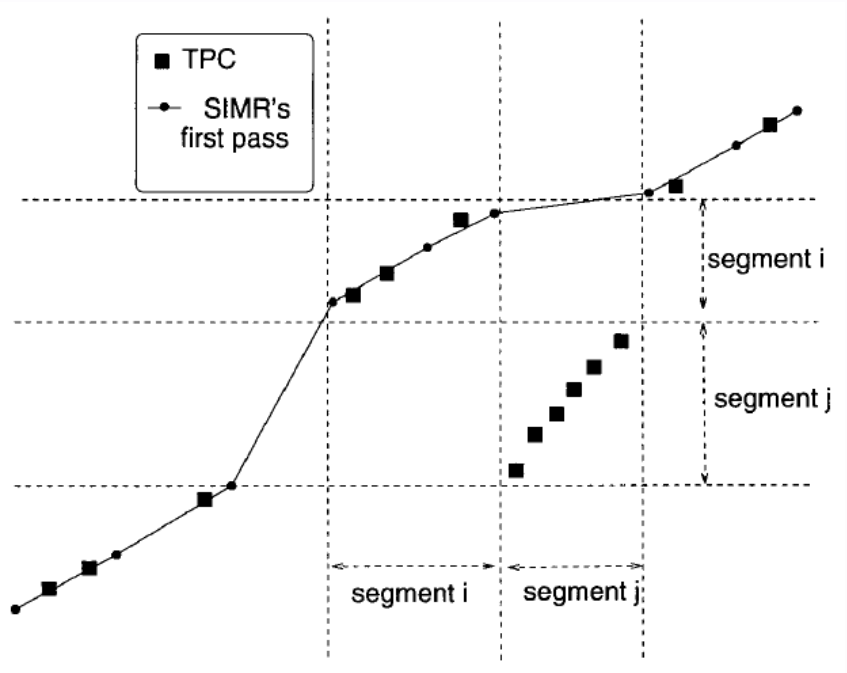




Passe supplémentaire



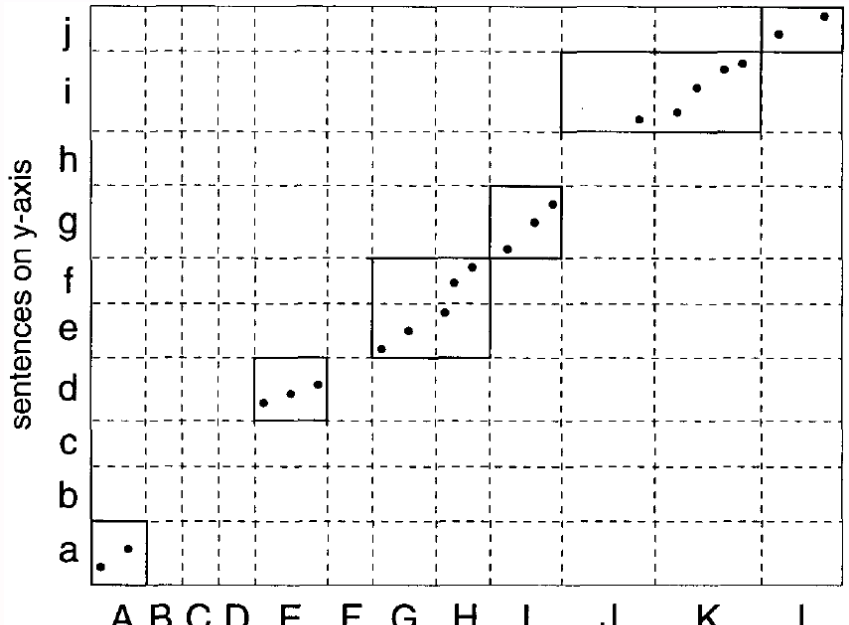
Prise en compte de la pente locale



Croisement de segments

Sélection des segments alignés : GSA

- S'appuie sur les chaînes de points de correspondance pour proposer des alignements de segments (phrases, paragraphes, listes, etc.)
- GSA : alignement géométrique de segments
- Information nécessaire : frontières de segments (p. ex., segmentation en phrases)
- Les segments doivent être contigus et ne pas se croiser



En deux passes (Moore, 2002)

(Moore, 2002)

Fonctionne en deux passes

- 1. Premier alignement selon la *longueur des phrases*
  - Sélection initiale de phrases alignées 1-1 avec une haute probabilité
  - $p > 99$ %, soit 80 % du corpus
- Construction d'un *modèle de traduction de mots*
- 2. Second alignement basé sur la longueur des phrases + le modèle de traduction

Méthode mixte qui limite le surcoût par rapport une méthode basée uniquement sur la longueur des phrases

`https://www.microsoft.com/en-us/download/details.aspx?id=52608`

En deux passes (Moore, 2002)

(Moore, 2002)

Fonctionne en deux passes

1. Premier alignement selon la *longueur des phrases*

en_1	Poor Alice!	Pauvre Alice !	fr_1
en_2	It was as much as she could do, lying down on one side, to look through into the garden with one eye; but to get through was more hopeless than ever : she sat down and began to cry again.	C'est tout ce qu'elle put faire, après s'être étendue de tout son long sur le côté, que de regarder du coin de l'oeil dans le jardin. Quant à traverser le passage, il n'y fallait plus songer. Elle s'assit donc, et se remit à pleurer.	fr_2 fr_3 fr_4
en_3	"You ought to be ashamed of yourself," said Alice,"a great girl like you," (she might well say this), "to go on crying in this way!	«Quelle honte !» dit Alice. «Une grande fille comme vous» («grande» était bien le mot) «pleurer de la sorte !	fr_5 fr_6
en_4	Stop this moment, I tell you!"	Allons, finissez, vous dis-je ! »	fr_7
en_5	But she went on all the same, shedding gallons of tears, until there was a large pool all round her, about four inches deep and reaching half down the hall.	Mais elle continue de pleurer, versant des torrents de larmes, si bien qu'elle se vit à la fin entourée d'une grande mare, profonde d'environ quatre pouces et s'étendant jusqu'au milieu de la salle.	fr_8

Table 6.3: An example alignment computed by Moore's algorithm for *Alice's Adventures in Wonderland*. The first and third anchor links delineate a 2×5 gap containing 2 English and 5 French sentences.

(Xu, 2016, p. 90)

Évaluation sur des textes littéraires

		GMA	BMA	Hun	Garg	Yasa
BAF		61.4	73.6	71.2	65.6	75.7
manual en-fr	min	53.5	57.4	54.3	51.7	59.9
	max	92.8	91.5	92.6	97.1	95.6
	mean	79.6	74.9	74.5	80.2	79.1
auto en-fr	min	62.1	47.1	56.6	56.4	62.3
	max	99.5	98.4	99.5	98.1	98.8
	mean	88.7	84.0	87.9	88.7	89.6
auto en-es	min	60.3	48.8	43.7	60.9	58.3
	max	96.5	98	96.4	98.8	98.4
	mean	82.8	78.4	81.0	80.5	82.7

Table 6.2: Baseline evaluation results

(Xu, 2016, p. 87)

Alignement par apprentissage supervisé

Détection de phrases parallèles dans des corpus comparables (Munteanu & Marcu, 2005)

- Paires de phrases candidates : produit cartésien des phrases source et cible
- Filtre initial
  - rapport des longueurs < 2
  - au moins la moitié des mots d'une phrase ont une traduction dans l'autre phrase
- Tâche : décider si une paire de phrases candidates sont ou pas traduction l'une de l'autre
- Exemples
  - positifs : paires de phrases parallèles
  - négatifs : paires de phrases qui ne sont pas parallèles
- Caractéristiques pour représenter chaque paire de phrases candidates

Caractéristiques

- Caractéristiques globales
  - longueur des deux phrases, différence et rapport des longueurs
  - pourcentage des mots de chaque phrase qui ont une traduction dans l'autre phrase (selon un dictionnaire bilingue)
- Caractéristiques liées à la qualité d'un alignement automatique des mots (qui connecte mots source et mots cible)
  - pourcentage et nombre des mots sans connexion
  - trois plus hautes « fertilités » (nombre de mots cible connectés à un mot source)
  - longueur du plus grand segment contigu connecté
  - longueur du plus grand segment non connecté

Alignement par apprentissage supervisé

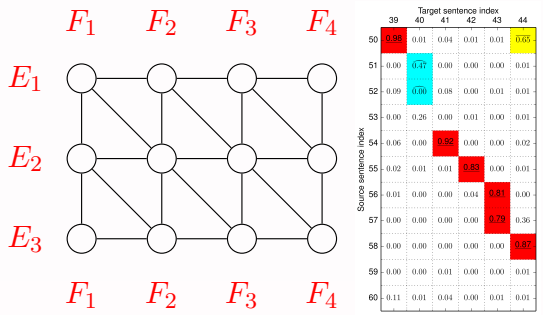
Alignement de phrases dans des textes parallèles (Yu *et al.*, 2012)

- Alignement en deux passes, comme (Moore, 2002)
- Deuxième passe : aligne les phrases entre les zones alignées
  - Paires de phrases candidates : produit cartésien des phrases source et cible de deux zones non alignées
- Apprentissage supervisé (régression logistique)
- Exemples
  - positifs : phrase source et sa phrase cible
  - négatifs : phrase source et la phrase suivant sa phrase cible
- Caractéristiques :
  - paires de mots (source, cible)
  - différence relative des longueurs des deux phrases : $\frac{|l_s - l_t|}{\max(l_s, l_t)}$

Autres exemples

Alignement de phrases dans des textes parallèles

- (Kaufmann, 2012) : Caractéristiques proches de (Munteanu & Marcu, 2005), régression logistique (MaxEnt)
- (Mújdricza-Maydt *et al.*, 2013) : champs conditionnels aléatoires (CRF)
  - Prend en compte le fait que la phrase précédente est alignée ou pas
- (Xu, 2016) : champs conditionnels aléatoires 2D (CRF)
  - Prend en compte davantage de dépendances entre alignements



(Xu, 2016, p. 94 ; 97)

  - Heuristiques pour lisser les alignements

Pour aller plus loin

- Bibliographie sur l'alignement de phrases :
  `http://www.statmt.org/survey/Topic/SentenceAlignment`
- États de l'art : (Wu, 2010; Tiedemann, 2011)

Bibliographie I


Gale W. & Church K. W. (1993).
A program for aligning sentences in bilingual corpora.
Computational Linguistics, **19**(3), 75–102.


Kaufmann M. (2012).
JMaxAlign : A maximum entropy parallel sentence alignment tool.
In *Proceedings of COLING 2012 : Demonstration Papers*, p. 277–288, Mumbai, India : The COLING 2012 Organizing Committee.


Melamed I. D. (1999).
Bitext maps and alignments via pattern recognition.
Computational Linguistics, **25**(1), 107–130.

Bibliographie II

Moore R. C. (2002).
Fast and accurate sentence alignment of bilingual corpora.
In *Machine Translation : From Research to Real Users*, p. 135–244, Heidelberg, Germany : Springer-Verlag.
Actes 5th Conference of the Association for Machine Translation in the Americas.

Munteanu D. S. & Marcu D. (2005).
Improving machine translation performance by exploiting non-parallel corpora.
Computational Linguistics, **31**(4), 477–504.

Májdricza-Maydt É., Körkel-Qu H., Riezler S. & Padó S. (2013).
High-precision sentence alignment by bootstrapping from wood standard annotations.
Prague Bulletin of Mathematical Linguistics, (99), 5–16.

Bibliographie III


Tiedemann J. (2011).
Bitext Alignment.
Synthesis Lectures on Human Language Technologies. Morgan & Claypool.


Véronis J. & Langlais P. (1999).
ARCADE : Évaluation de systèmes d'alignement de textes multilingues.
In *Actes JST 99*.


Wu D. (2010).
Alignment.
In *Handbook of Natural Language Processing, Second Edition*, p. 367–408. Chapman and Hall/CRC.


Xu Y. (2016).
Confidence Measures for Alignment and for Machine Translation.
Thèse de doctorat en informatique, Université Paris-Saclay, Orsay.

Bibliographie IV

Yu Q., Max A. & Yvon F. (2012).
Revisiting sentence alignment algorithms for alignment visualization and evaluation.
In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, p. 10–16, Istambul, Turkey.