Modes d'évaluation et effets de mode dans l'évaluation de la Traduction Automatique:

Que peut-on attendre des métriques d'évaluation de la Traduction Automatique ?

A. Balvet

02/02/2018

Traduction & Qualité: Biotraduction et Traduction Automatique

Université de Lille

STL

Linguistique

Linguists love ambiguity more than most people

Linguistique

Linguists love ambiguity more than most people

Google : Les linguistes aiment l'ambiguïté plus que la plupart des gens ✓

Linguistique

Linguists love ambiguity more than most people

Google : Les linguistes aiment l'ambiguïté plus que la plupart des gens ✓
Systran : Ambiguïté d'amour de linguistes plus que la plupart des personnes ✗

Machine Translation is here to stay !

Machine Translation is here to stay !

Google/Systran : la traduction automatique est là/ici pour rester ✗

ALPAC

Language and Machines, Computers in Translation and Linguistics. Automatic Language Processing Advisory Committee, National Academy of Sciences National Research Council, 1966

ALPAC

QUALITY

The Committee believes strongly that the quality of translation must be adequate to the needs of the requester. (...)

Despite the fact that adequate quality is essential, ***the government has no reliable way to measure the quality of translation****. In view of this, one member of the Committee has set up an experiment in the evaluation of quality. (...) A reliable way to measure quality would be of great importance in determining proper cost of translation. The correlation between cost and quality is far from precise.*

[Pierce and Carroll, 1966, p. 16]

ALPAC

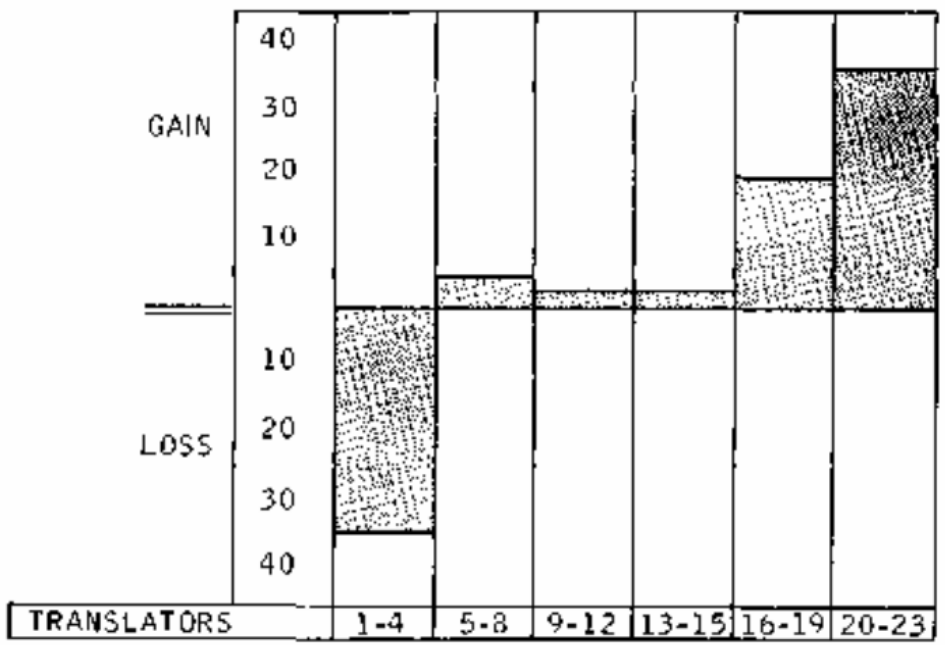


FIGURE 3. Percentage gain or loss in output from postediting.

Plan

Aperçu historique

Aperçu historique

Aperçu historique

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

ALPAC

ESTIMATES OF COSTS AND QUALITY FOR VARIOUS TYPES OF TRANSLATION

Type	Quality	Cost for 11,000,000 Russian Words
In-house (conventional translation)	Good	$ 440,000
In-house (dictation)	Good	440,000-
Contract	Fair to good	350,000
JPRS	Fair	240,000
Raw MT	Unsatisfactory	80,000 +
Postedited MT	Fair	400,000
M-AT	Excellent	310,000
Analysts proficient in Russian	-	0

[Pierce and Carroll, 1966]

Tendances

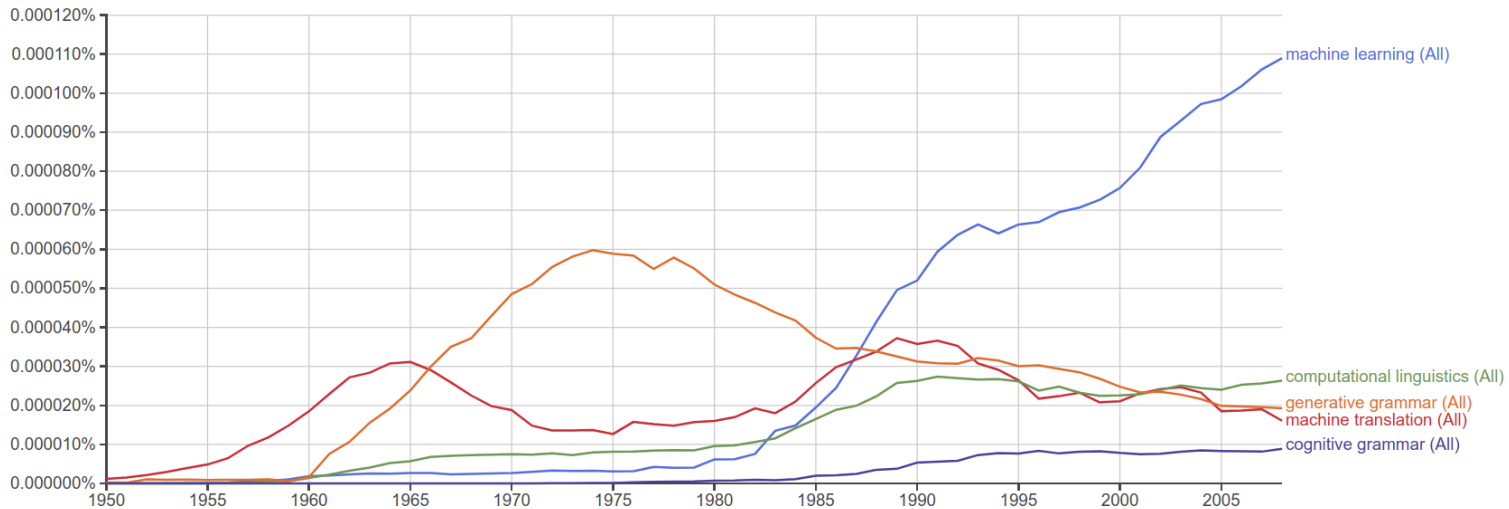
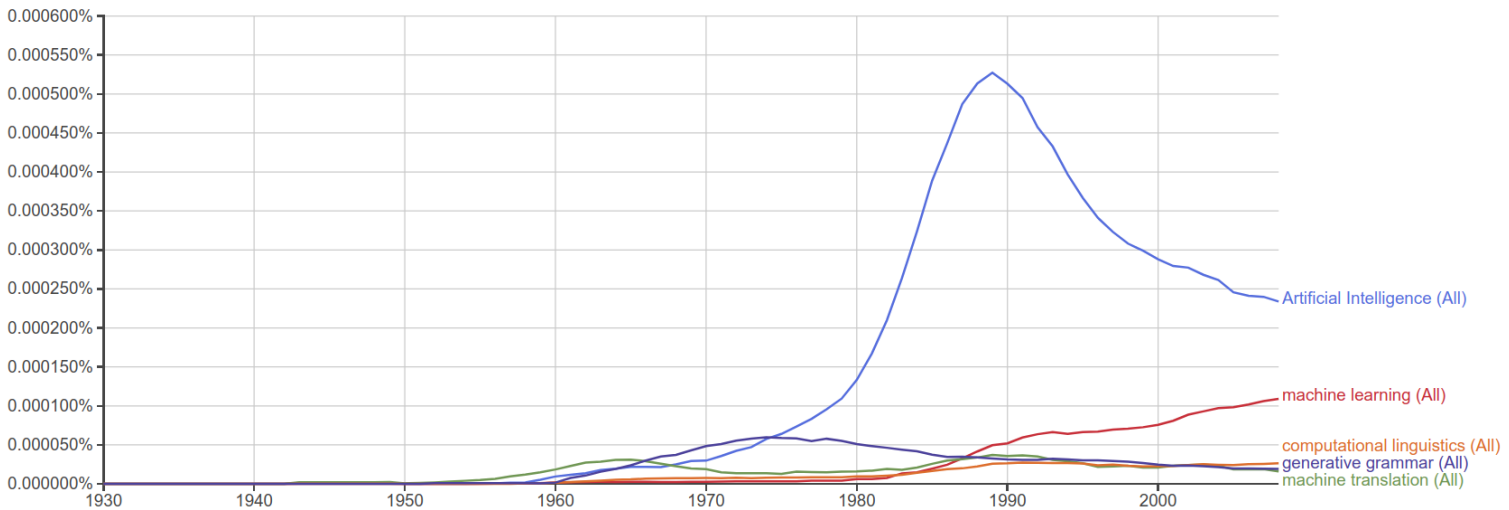
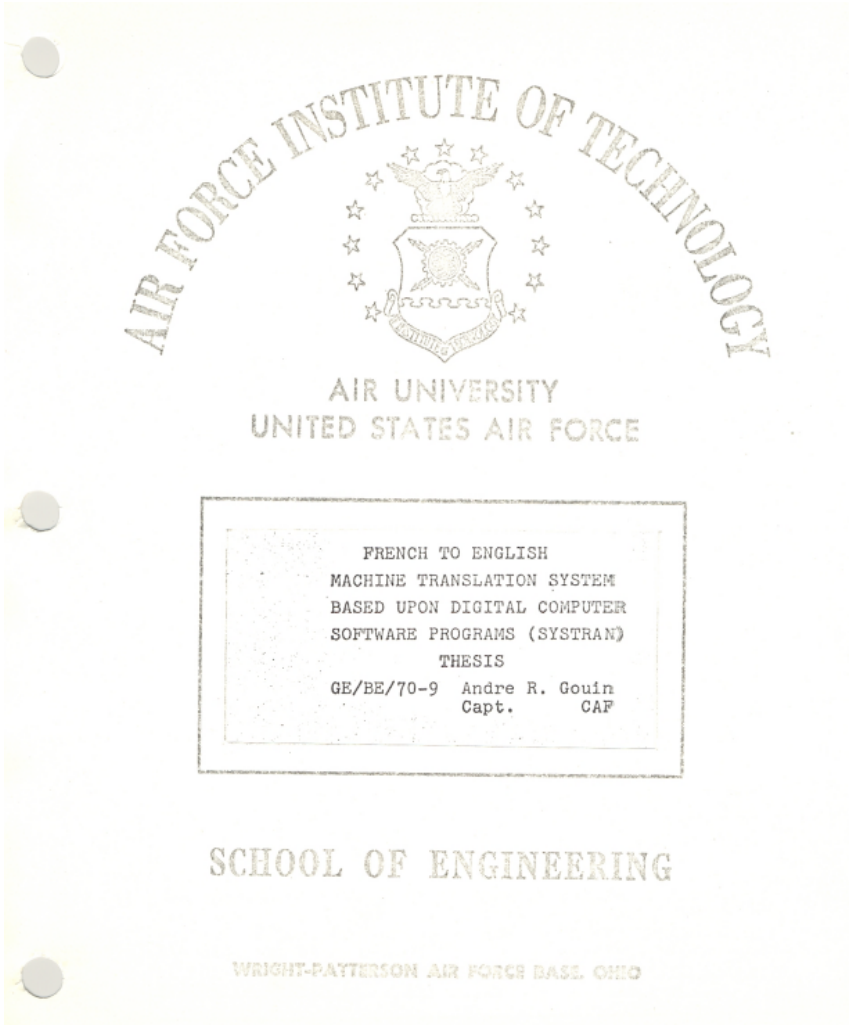


Figure 1 – Tendances : ML/NLP (Google ngrams)

Tendances : AI, ML, MT...



Premiers systèmes industriels : Systran (1970)



AIR FORCE INSTITUTE OF TECHNOLOGY

AIR UNIVERSITY
UNITED STATES AIR FORCE

FRENCH TO ENGLISH
MACHINE TRANSLATION SYSTEM
BASED UPON DIGITAL COMPUTER
SOFTWARE PROGRAMS (SYSTRAN)
THESIS

GE/BE/70-9 Andre R. Gouin
Capt. CAF

SCHOOL OF ENGINEERING

WRIGHT-PATTERSON AIR FORCE BASE, OHIO

Systran II (1975)

De la Guerre Froide à la coopération scientifique

Breaking the Language Barrier

Heading a group of technology transfers in the field of computer processing is a space-spurred electronic translation system

De la Guerre Froide à la coopération scientifique

The key element of SYSTRAN II is a computer program—one of the longest ever written, with half a million lines of instructions—backed by a computerized dictionary which contains terminology, technical expressions, grammatical rules and semantic principles. The text to be translated is fed into the computer, which analyzes it for syntax and semantics, then produces—in printout form—an accurate version of the text in the target language. The computer's draft is refined by human translators, whose editing is also computerized. The system then produces a magnetic tape ready for photocomposition.

Des systèmes à base de règles...

MYCIN

RULE037

IF: 1) *The identity of the organism is not known with* certainty, and
2) The stain of the organism is gramneg, and
3) The morphology of the organism is rod, and
4) The aerobicity of the organism is aerobic

THEN: There is strongly suggestive evidence (.8) that the class of the organism is enterobacteriaceae

RULE145

IF: 1) The therapy under consideration is one of: cephalothin clindamycin erythromycin lincomycin vancomycin, and
2) Meningitis is an infectious disease diagnosis *for the patient*

THEN: It is definite (1) the therapy under consideration is not a potential therapy for use against the organism

RULE060

IF: *The identity of the organism is bacteroides*

THEN: I recommend therapy chosen from among the following drugs:

1 - clindamycin	(.99)
2 - chloramphenicol	(.99)
3 - erythromycin	(.57)
4 - tetracycline	(.28)
5 - carbenicillin	(.27)

[Buchanan and Shortliffe, 1984]

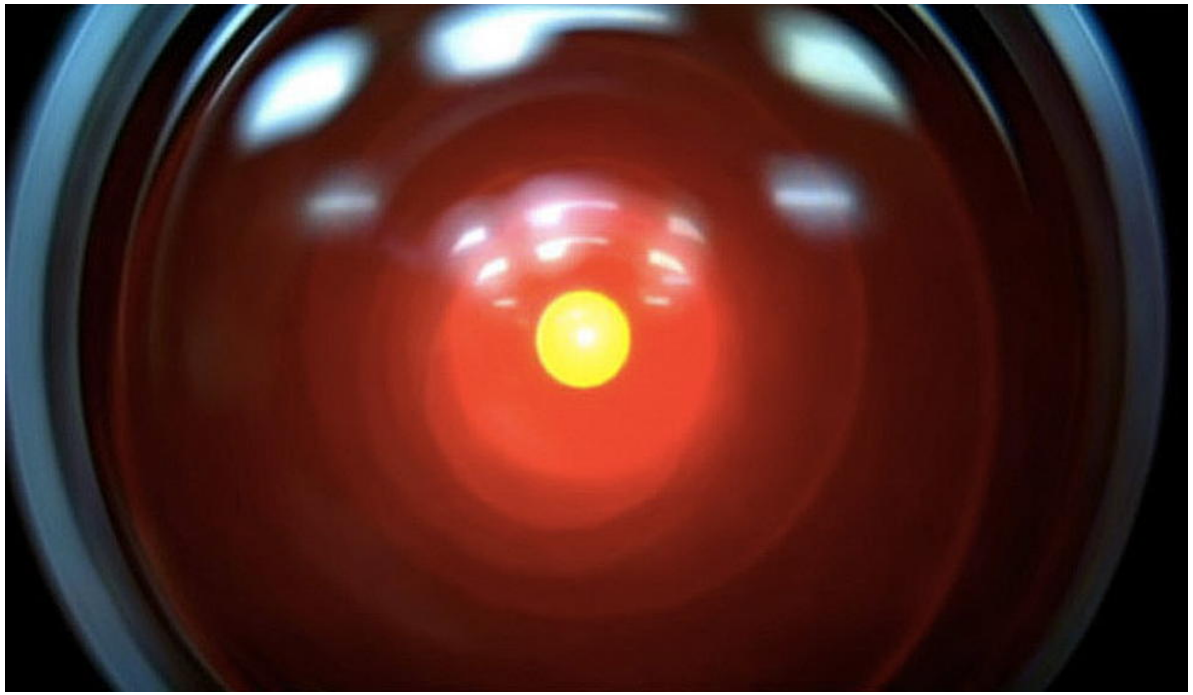
... aux modèles de traduction paramétrés automatiquement

Induction de candidats traductions à partir de corpus parallèles

```
yo soy -> ['that's me', 'that's', 'i am']
estoy -> ['i', 'am', 'am i']
estoy yo -> ['am i', 'am i not', 'am']
qué pasa -> ['what's happening', 'what's', 'happening']
es su casa -> ['is his house', 'is his', 'this is his']
quiero hablar con -> ['i wish to', 'i wish', 'wish to speak
vete -> ['go', 'go in', 'in']
bebamos -> ['lets', 'lets drink', 'drink']
por qué -> ['why', 'for what', 'for']
la mesa -> ['the', 'table', 'the table']
en la mesa -> ['at the table', 'at the', 'part at the']
```

La TA aujourd'hui

La TA aujourd'hui

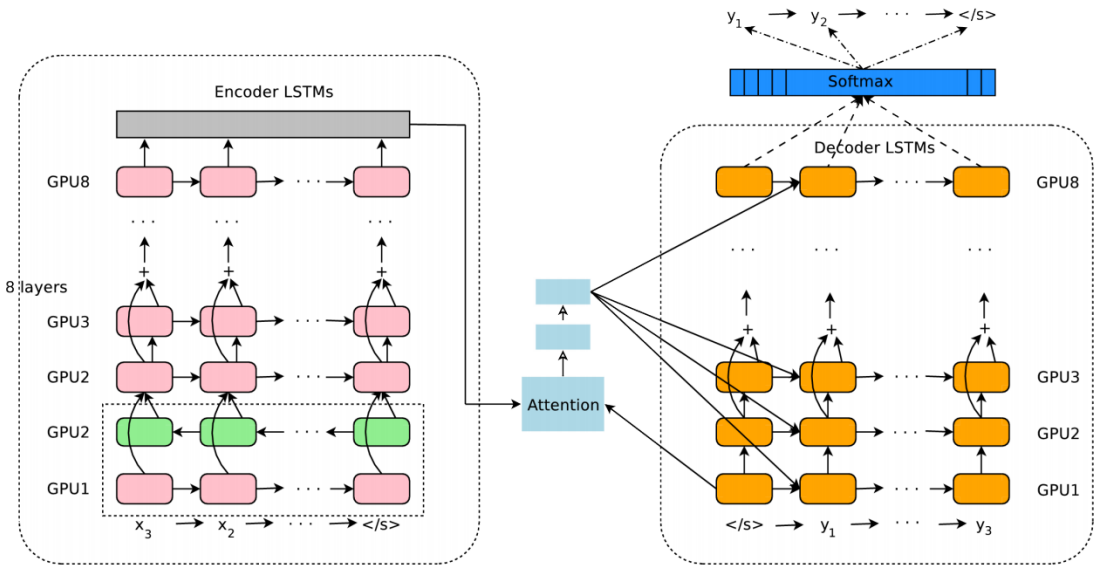


La TA aujourd'hui

Your Google Assistant,
always ready to help.

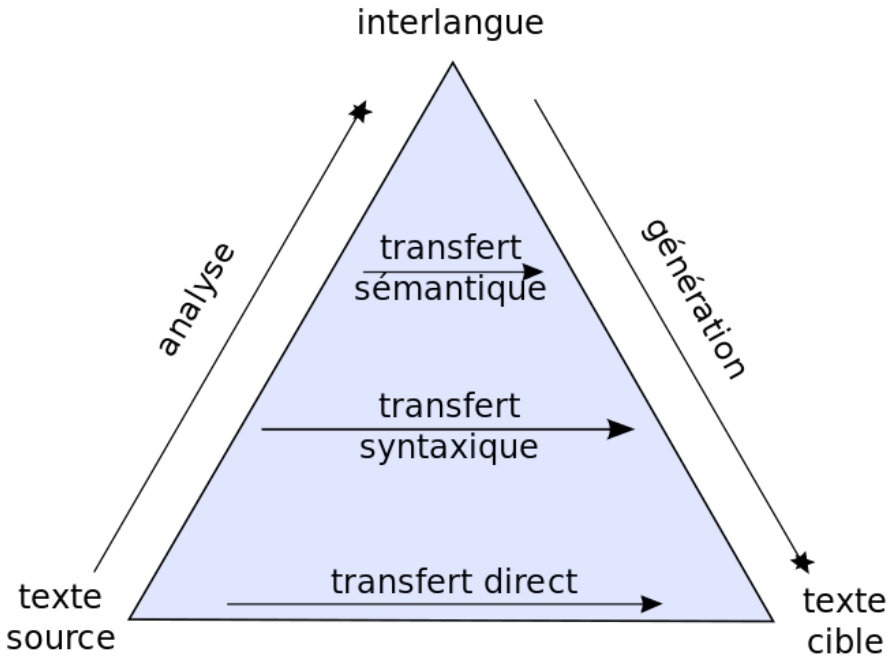
La TA aujourd'hui

Google NMT


Encoder LSTMs
GPU8
8 layers
GPU3
GPU2
GPU2
GPU1
x3 x2 ... </s>
Attention
y1 y2 ... </s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s> y1 ... y3

Principales approches

Principales approches



Triangle de Vauquois

Principales approches

Aujourd'hui, 2 grandes familles :

1. "Rule-Based MT" (RBMT)
   - dictionnaires bilingues
   - règles d'analyse et de transformation
2. Data-Driven : induction d'un modèle de transfert langue source / langue cible
   - "Example-based MT" (EBMT) [Nagao, 1984] (traduction par analogie)
   - "Statistical MT" (SMT) [Brown et al., 1988]
   - "Phrase-Based MT"
   - "Neural MT"

Principales approches

Approches statistiques, évoquées dans [Weaver, 1955] et réintroduites par les chercheurs d'IBM Watson : [Brown et al., 1988]

- "Statistical MT" (SMT) :
  - estimer au mieux la distribution de probabilité $p(e|f)$ qu'une chaîne de caractère e dans la langue cible soit la traduction d'une autre chaîne f dans la langue source
  - passage de traduction "word-based" à "phrase-based" (PBSMT) et au-delà

$\Rightarrow$ distinction entre TA guidée par l'expertise humaine *vs.* induction à partir de corpus parallèles alignés

SMT

Probabilités lexicales induites à partir d'un corpus parallèle, [Koehn, 2009]

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

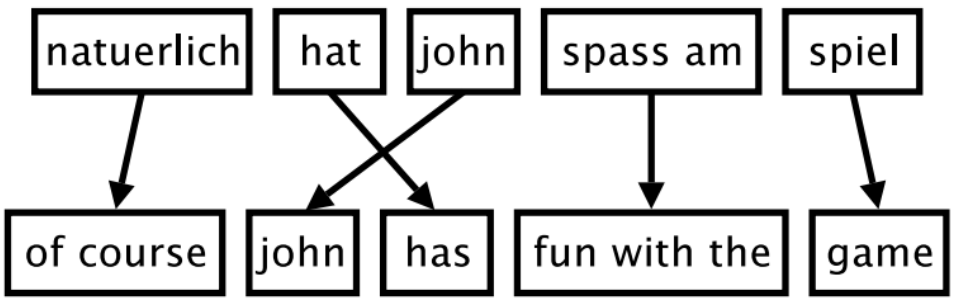
e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

PBSMT

Exemple d'alignement de "phrases", [Koehn, 2009]



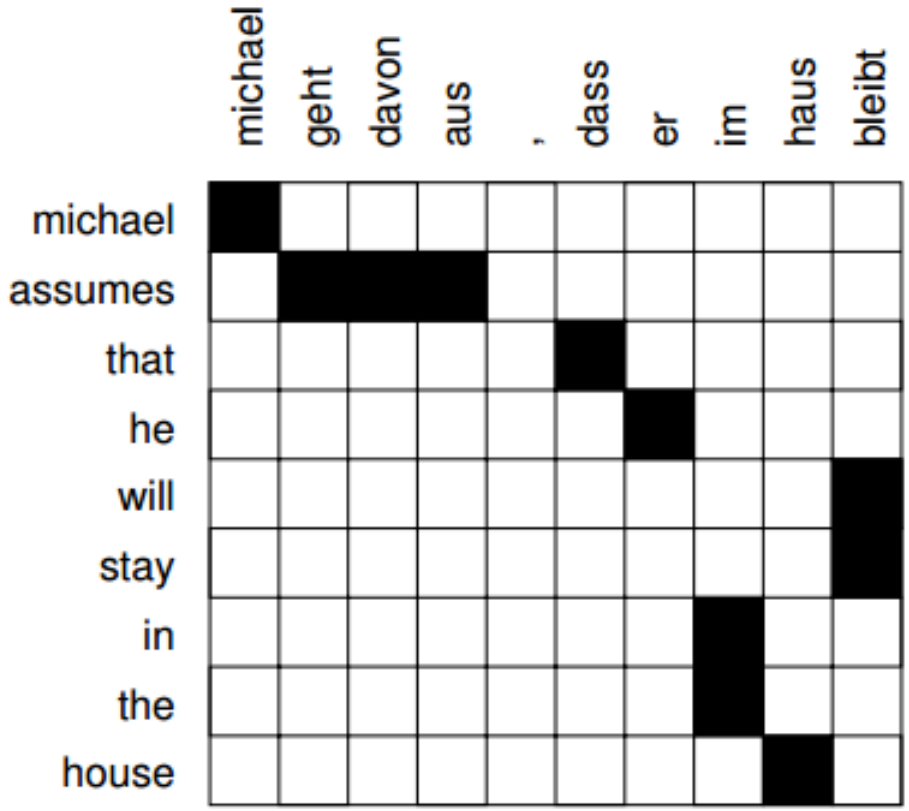
PBSMT

Exemple de “phrases” induites à partir du corpus Europarl, pour la séquence “den Vorschlag”, [Koehn, 2009]

English	$\phi(\bar{e}\|\bar{f})$	**English**	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
’s proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

PBSMT

Alignement de mots, [Koehn, 2009]



NMT

Application du Machine Learning (neurones artificiels) à la TA

- premiers réseaux de neurones artificiels : [McCulloch and Pitts, 1943], [Rosenblatt, 1958]
- premier ouvrage théorique [Hebb, 1949]
- premier système de NMT “deep learning” : [Sutskever et al., 2014]

Tendances

Figure 2 – Tendances en machine learning (Google ngrams)

Tendances

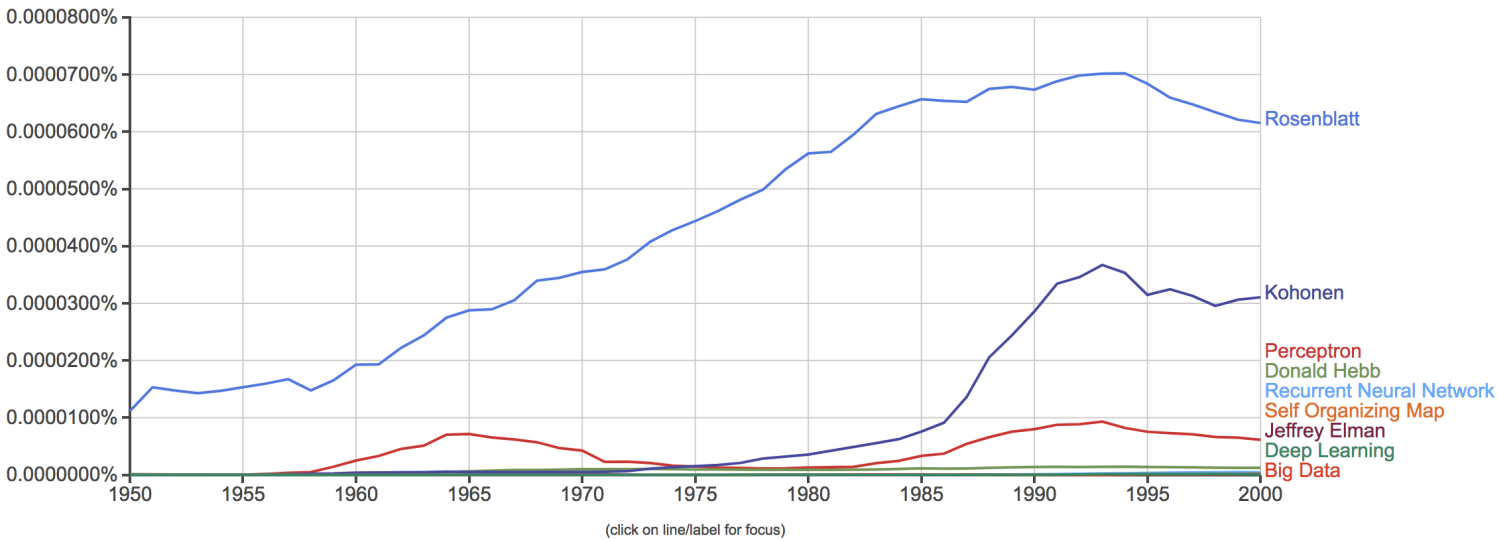
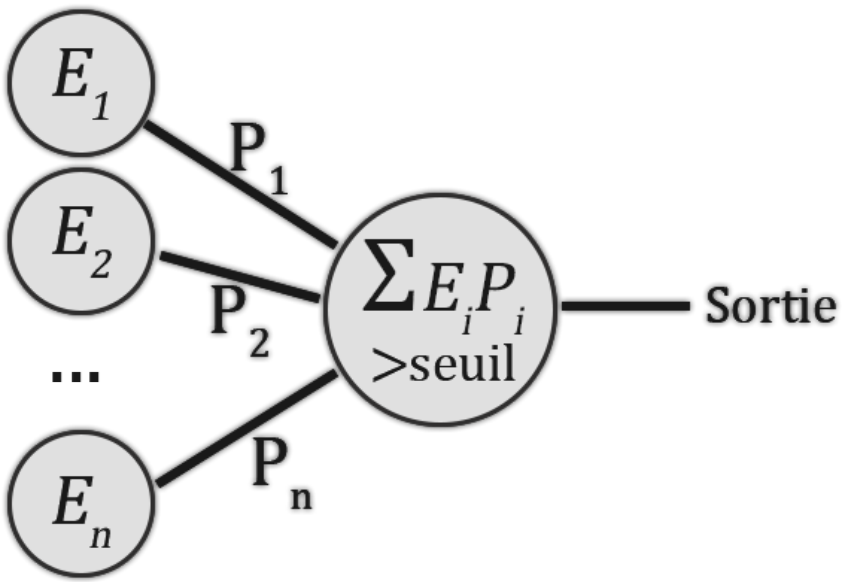


Figure 3 – Tendances : ML/linguistique (Google ngrams)

ANN



Perceptron

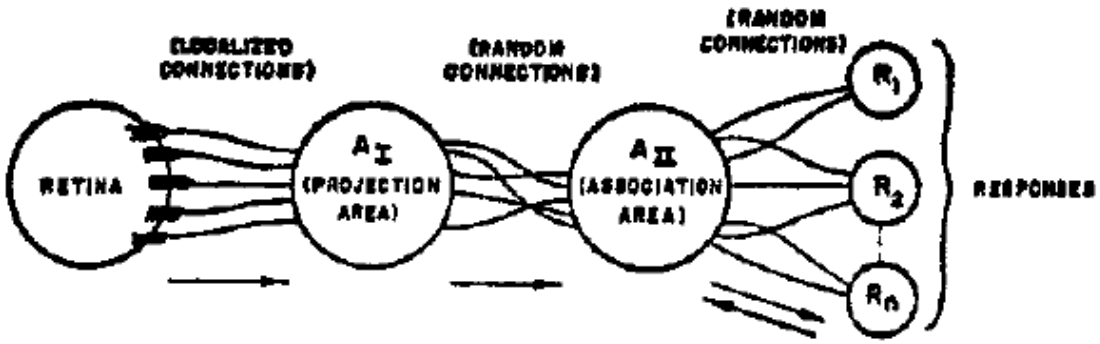
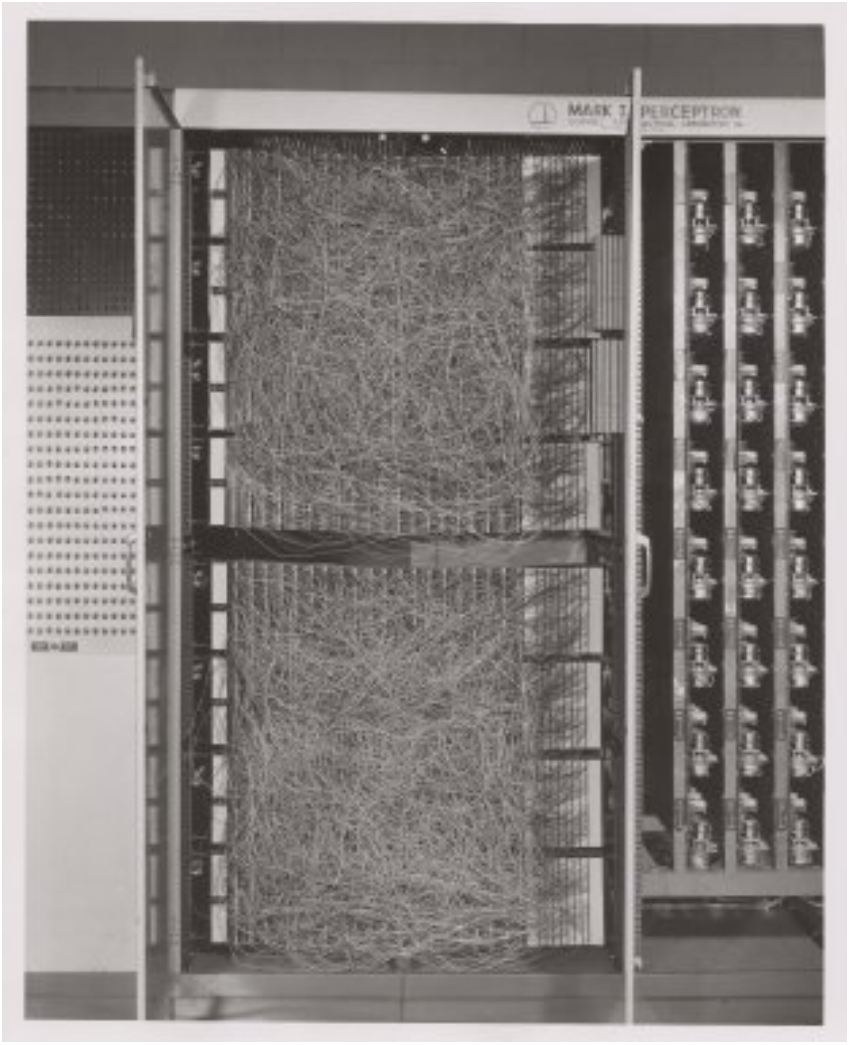


FIG. 1. Organization of a perceptron.

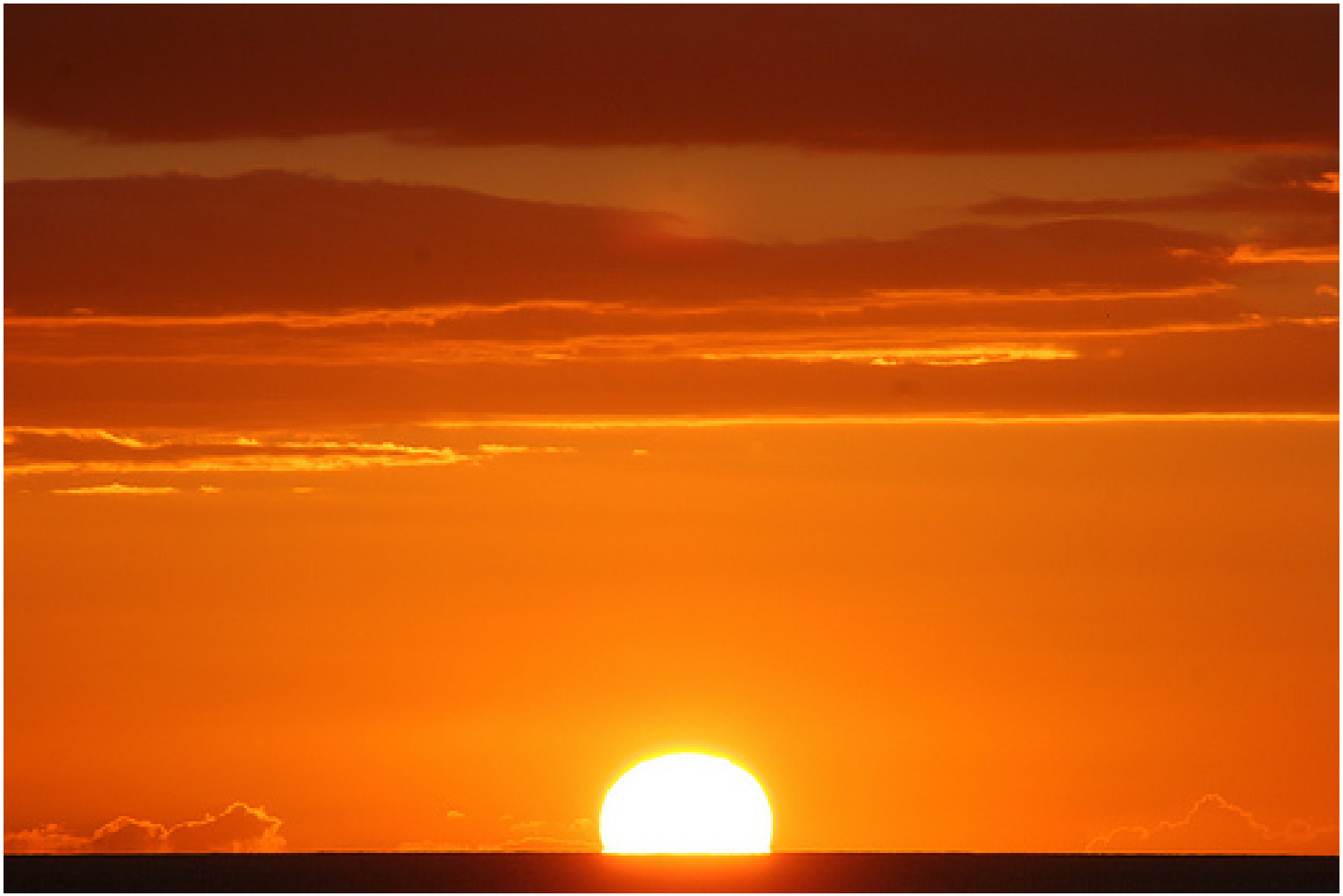
Perceptron, Mark I



NMT

Les GAFAM se positionnent sur ce marché :

- mise à disposition de projets de NMT en open-source
- Facebook : https ://github.com/facebookresearch/fairseq-py (Sequence-to-sequence MT)
- Google : https ://github.com/tensorflow
  - Syntaxnet
  - Google NMT



Plus. Jamais. ÇA ! (?)

Ou ÇA !

→ “干” (gàn) = “sec” ou ...

Ni ÇA !


妇产科
Gynecology & Obstetrics Dept.
抢救室 Emergency Room
阴道镜检查室 Cunt Examination
胎心监护室 Fetal Heart Custody
药流观察室 Observation Room
产科诊室 Obstetrics Consultation

Et encore moins ÇA !

Ou ÇA !

Bandi

Bandi per assegni di ricerca

Univ. FIRENZE

Bando per assegno di ricerca

Descrizione del bando

Titolo del progetto di ricerca in italiano	'Dalla pecora al pecorino' tracciabilità e rintracciabilità di filiera nel settore lattiero caseario toscano
Titolo del progetto di ricerca in inglese	'From sheep to Doggy Style' traceability of milk chain in Tuscany
Campo principale della ricerca	Agricultural sciences
Sottocampo della ricerca	Zootechnics
Area CUN	07 – Scienze agrarie e veterinarie
S.S.D	AGR/19 – ZOOTECNICA SPECIALE
Descrizione sintetica in italiano	Il progetto rappresenta l'implementazione a livello territoriale di uno studio già realizzato a livello di un singolo caseificio e dei produttori conferenti, si sviluppa attraverso l'impiego di componenti di rilevazione e registrazione degli eventi, applicate ad una metodologia che utilizza hardware e software tra loro integrati. Ogni giro di raccolta è univocamente identificato attraverso la messa a punto di un protocollo di comunicazione basato sulla tecnologia GPS, con possibilità di riconoscimento dei tank refrigeranti delle aziende di produzione, delle autocisterne e dei tank presso il caseificio. La produzione di latte è controllata al momento della raccolta, mediante misurazione magneto-induttiva della quantità di latte raccolto; la rintracciabilità del latte prosegue durante il giro di raccolta attraverso l'identificazione degli scomparti dell'autocisterna fino al conferimento al caseificio.

TA : quelle qualité ?

Principes généraux

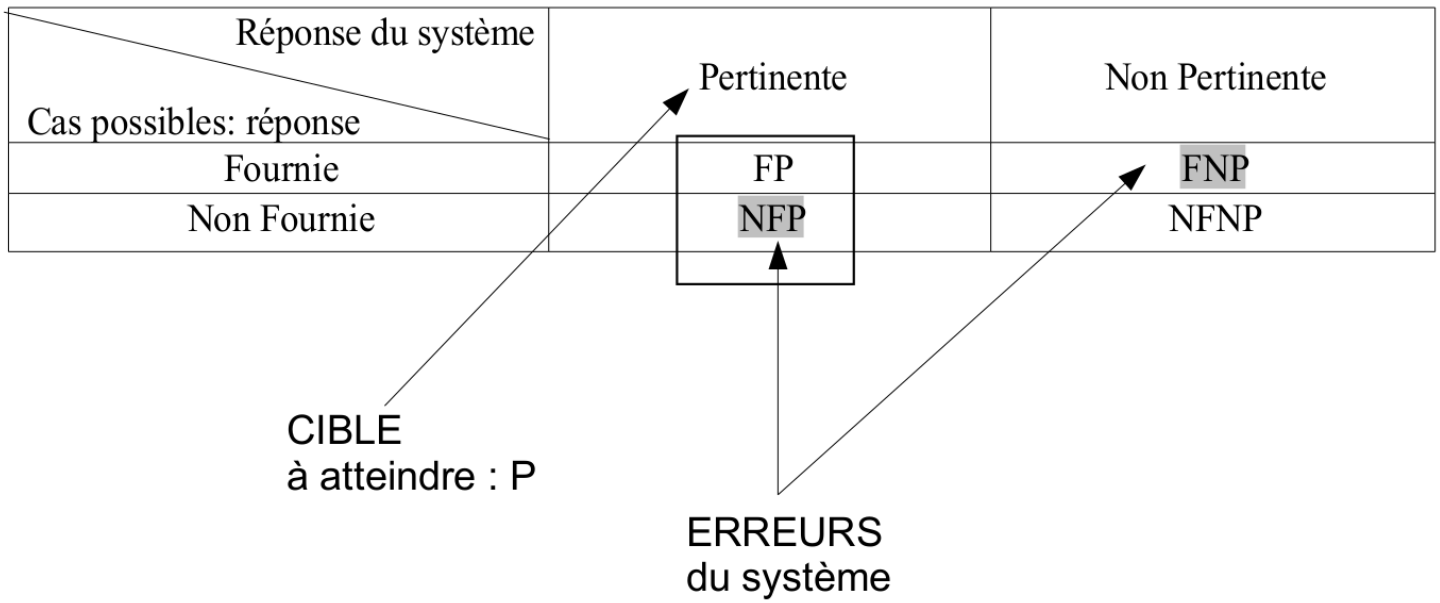
Quel que soit le domaine (IA, TAL, RI), il est nécessaire de disposer de métriques "objectives" pour évaluer les systèmes
Deux approches : évaluation extrinsèque (référence) / intrinsèque

- évaluation extrinsèque par référence, ou "gold standard"
  - la plupart du temps, référence constituée de façon (semi) manuelle
  - ex. : corpus de traduction de référence

En TA, plusieurs métriques ont été élaborées, dérivées en partie du domaine de la génération/résumé automatiques de texte

Précision et Rappel

La base de (presque) toutes les métriques en évaluation de systèmes de TAL/RI



Précision et Rappel

Formules de calcul

- Recall (rappel) : $\frac{FP}{FP+NFP}$
  - proportion de FP par rapport à la cible à atteindre (P)
- Precision : $\frac{FP}{FP+FNP}$
  - proportion de P parmi les réponses fournies (F)

Principes généraux

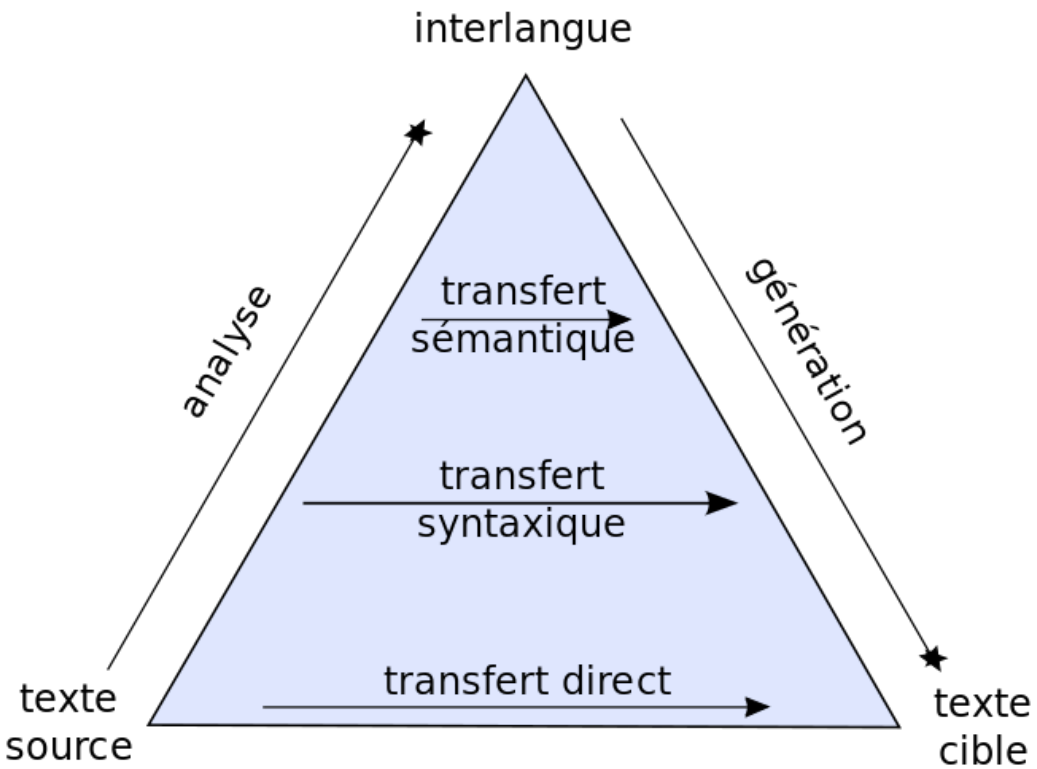
Approche extrinsèque : la nécessité d'une référence

- corpus de traductions validées par des experts humains
- sorties des systèmes de TA évaluées par des experts humains

$\Rightarrow$ dépendance envers l'expertise humaine
$\Rightarrow$ difficultés liées à la variabilité des jugements (inter- et intra-personnelle)

$\Rightarrow$ évaluation intrinsèque ?

Principes généraux



ALPAC : comment évaluer la qualité de la TA ?

THE MEASUREMENT PROCEDURE

It was reasoned that the two major characteristics of a translation are (a) its intelligibility, and (b) its fidelity to the sense of the original text. Conceptually, these characteristics are independent ; that is, a translation could be highly intelligible and yet lacking in fidelity or accuracy. Conversely, a translation could be highly accurate and yet lacking in intelligibility; this would be likely to occur, however, only in cases where the original had low intelligibility.

ALPAC : échelle d'intelligibilité

TABLE 4. Scale of Intelligibility

9—Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.

8—Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected."

7—Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.

6—The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in nearly acceptable form.

5—The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.

4—Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.

3—Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.

2—Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.

1—Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

Principes généraux

Problème :

- aucun système de TA ne peut espérer fournir un output qui soit exactement celui de la référence
- aucun traducteur humain ne traduit exactement comme la référence
  - ex. : choix de la voix passive vs. active
  - ex. : choix d'une nominalisation vs. verbe
  - ex. : phrases longues / courtes

$\Rightarrow$ problème général de la variabilité des formes associées à un contenu informationnel/sens

Choix de traduction

Fidélité relative par rapport à la source :

- "To be or not to be, that is the question !" (Hamlet, W. Shakespeare)
  1. "Être ou ne pas être, telle est la question !" (Google)
  2. "Être ou ne pas être, c'est la question !" (Systran, Morand & Schwob)

Choix de traduction

Fidélité relative par rapport à la source :

- "To be or not to be, that is the question !" (Hamlet, W. Shakespeare)
  1. "Être ou ne pas être, telle est la question !" (Google)
  2. "Être ou ne pas être, c'est la question !" (Systran, Morand & Schwob)
  3. "Être ou n'être pas, voilà tout le problème" (Ménard, 1866)
  4. "Être ou ne pas être, tout est là !" (S. Becker, TNP Villeurbanne)
  5. "Vivre encore ou cesser de vivre ? Voilà bien ce qu'il faut choisir" (R. Y. C. Mauroy)

Choix de traduction

Fidélité relative par rapport à la source :

- "To be or not to be, that is the question !" (Hamlet, W. Shakespeare)
  1. "Être ou ne pas être, telle est la question !" (Google)
  2. "Être ou ne pas être, c'est la question !" (Systran, Morand & Schwob)
  3. "Être ou n'être pas, voilà tout le problème" (Ménard, 1866)
  4. "Être ou ne pas être, tout est là !" (S. Becker, TNP Villeurbanne)
  5. "Vivre encore ou cesser de vivre ? Voilà bien ce qu'il faut choisir" (R. Y. C. Mauroy)
  6. "Demeure ; il faut choisir, et passer à l'instant
     De la vie à la mort, et de l'être au néant" (Voltaire, 1733)

Principes généraux

Réponse :

- l'évaluation ne porte pas sur le segment "phrase" mais sur un sous-ensemble normalisé de ce segment : des n-grammes
- on distingue entre correspondance exacte et partielle (similarité)
  - choix des items lexicaux
  - ordonnancement

Un système de TA peut avoir

- tous les "bons" mots, mais pas dans le bon ordre
- une partie seulement des mots-cible, mais dans le bon ordre

n-grammes

Soit un texte

- A B C D E F G H ... Z

Les n-grammes correspondants sont :

- n = 1 : [A, B, C, D ... , Z]
- n = 2 : [A B, B C, C D, D E, ... , Y Z]
- n = 3 : [A B C, B C D, C D E, D E F, ... , X Y Z]

BLEU

BLEU (Bilingual Evaluation Understudy), IBM, [Papineni et al., 2002]

- recherche de la meilleure correspondance entre sorties du système et référence humaine
- comparaison **segments** traduits/références humaines
- variante de la Précision
- pondération de tous les n-grammes ($n = 1...4$) segment par segment, score moyen calculé **sur tout le corpus**
- scores BLEU définis sur [0, 1]
  - 0 = correspondance nulle entre n-grammes de référence et n-grammes fournis par le système
  - 1 = correspondance parfaite
  - dans la pratique, même une traduction humaine n'atteindra jamais 1

BLEU

Métrique robuste, qui s'est imposée dans le domaine malgré les critiques

- score valable pour l'ensemble du corpus, pas au niveau du segment/phrase
- centrée sur la correspondance entre output et référence au niveau du mot ou du n-gramme
- ne renseigne pas suffisamment sur la lisibilité ou de la bonne formation des segments produits par le système
- favorise les traductions courtes
- estimer l'équivalent du Rappel fait appel à des métriques complexes et n'évite pas tous les écueils
- dépendante d'une bonne segmentation en mots
  - pas adaptée pour les langues "sans" indices explicites de frontières de mots : agglutinantes, voire asiatiques (Thaï)
- une augmentation du score BLEU n'est pas forcément corrélée à une augmentation de la qualité de la traduction

The ugly truth about BLEU

8.2 Evaluation Metrics

We evaluate our models using the standard BLEU score metric. To be comparable to previous work [41, 31, 45], we report tokenized BLEU score as computed by the `multi-bleu.pl` script, downloaded from the public implementation of Moses (on Github), which is also used in [31].

As is well-known, BLEU score does not fully capture the quality of a translation. For that reason we also carry out side-by-side (SxS) evaluations where we have human raters evaluate and compare the quality of two translations presented side by side for a given source sentence. Side-by-side scores range from 0 to 6, with a score of 0 meaning *"completely nonsense translation"*, and a score of 6 meaning *"perfect translation: the meaning of the translation is completely consistent with the source, and the grammar is correct"*. A translation is given a score of 4 if *"the sentence retains most of the meaning of the source sentence, but may have some grammar mistakes"*, and a translation is given a score of 2 if *"the sentence preserves some of the meaning of the source sentence but misses significant parts"*. These scores are generated by human raters who are fluent in both languages and hence often capture translation quality better than BLEU scores.

[Wu et al., 2016]

ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- métrique + logiciels de calcul, proposé par [Lin, 2004] et [Lin and Och, 2004]
  - au départ pensée pour l'évaluation en résumé automatique
- proportion de n-grammes faisant partie des segments de référence / n-grammes produits par le système
- différentes variantes : ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, ROUGE-SU, etc.
  - n-grammes de taille n les plus fréquents, plus longue sous-chaîne commune (LCS), pondérée (W) ou non (L), avec "trous" (S) etc.

ROUGE et LCS

X : [**A** **B** **C** **D** E F G]

ROUGE et LCS

X : [**A** **B** **C** **D** E F G]

Y_1 : [**A** **B** **C** **D** H I K]

ROUGE et LCS

X : [**A** **B** **C** **D** E F G]

Y_1 : [A B C D H I K]
Y_2 : [A H B K C I D]

$\Rightarrow$ $Y_1 et Y_2$ ont le même nombre de correspondances avec la cible X ([A, B, C, D])
mais Y_1 est plus **fidèle** à la cible : les segments de Y_1 sont (partiellement) alignés avec ceux de la cible

ROUGE

Bonne corrélation avec des jugements humains (Document Understanding Conference 2003 : résumés très courts)

	DUC 2003 10 WORDS SINGLE DOC					
	1 REF	4REFS	1 REF	4 REFS	1 REF	4 REFS
Method	CASE		STEM		STOP	
R-1	0.96	0.95	0.95	0.95	0.90	0.90
R-2	0.75	0.76	0.75	0.75	0.76	0.77
R-3	0.71	0.70	0.70	0.68	0.73	0.70
R-4	0.64	0.65	0.62	0.63	0.69	0.66
R-5	0.62	0.64	0.60	0.63	0.63	0.60
R-6	0.57	0.62	0.55	0.61	0.46	0.54
R-7	0.56	0.56	0.58	0.60	0.46	0.44
R-8	0.55	0.53	0.54	0.55	0.00	0.24
R-9	0.51	0.47	0.51	0.49	0.00	0.14
R-L	0.97	0.96	0.97	0.96	0.97	0.96
R-S*	0.89	0.87	0.88	0.85	0.95	0.92
R-S4	0.88	0.89	0.88	0.88	0.95	0.96
R-S9	0.92	0.92	0.92	0.91	0.97	0.95
R-SU*	0.93	0.90	0.91	0.89	0.96	0.94
R-SU4	0.97	0.96	0.96	0.95	0.98	0.97
R-SU9	0.97	0.95	0.96	0.94	0.97	0.95
R-W-1.2	0.96	0.96	0.96	0.96	0.96	0.96

METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering)

- moyenne harmonique du score de Précision et Rappel pour les 1-grammes (= mots)
- recherche des alignements entre les segments de l'output et ceux de la cible
  - sélection des correspondances préservant l'ordre des mots output / cible
  - priorité au Rappel sur la Précision
  - pénalité pour des n-grammes non adjacents
  - mot exact, variante(s) morphologique(s), synonymes validés
  - comparaison possible entre un output et plusieurs références

METEOR

Recherche d'une bonne corrélation avec le jugement humain au niveau des segments (phrase) plutôt que sur l'ensemble du texte produit, [Banerjee and Lavie, 2005]

- corrélations avec le jugement humain au niveau du corpus entier :
  - METEOR (0.964) > BLEU (0.817)
- corrélation avec le jugement humain au niveau de la phrase :
  - METEOR (0.403) > BLEU ✗

METEOR

X : the cat sat on the mat

Y_1	on the mat sat the cat	Score : 0.5000
Y_2	the cat sat on the mat	Score : 0.9977
Y_3	the cat **was** sat on the mat	Score : 0.9654

NIST

NIST (National Institute of Standards and Technology), [Doddington, 2002]

- adaptation de BLEU
  - Précision au niveau du n-gramme + pondération positive des correspondances sur des n-grammes peu fréquents dans la référence
  - ex. : $w(\textit{interesting}, \textit{calculations}) > w(\textit{on}, \textit{the})$
- NIST présenterait une meilleure adéquation avec les jugements humains que BLEU

WER

WER (Word Error Rate)

- formule de calcul basée sur la correspondance explicite, mot exact ou variante morphologique
- d'habitude plutôt utilisée en reconnaissance vocale
  - calcul de similarité (Levenshtein) entre l'output et la cible, au niveau du mot
- permet de comparer plusieurs systèmes entre eux, ou l'évolution de performance d'un système donné
- formule : $WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$
  - S = Substitutions, D = Deletions, I = Insertions
  - C = Correct (correspondance exacte)
  - possibilité de pondérer les opérations d'édition (S, I, D) indépendamment

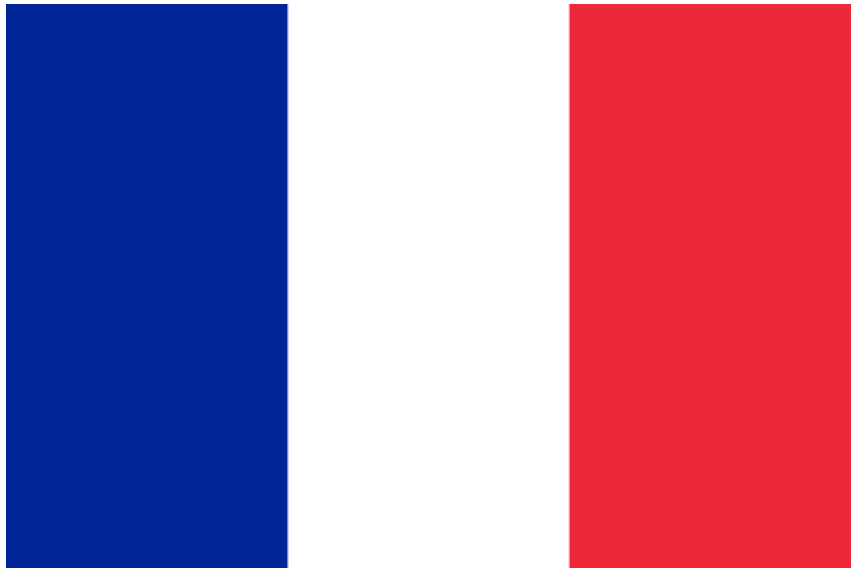
Métriques composites ?

Sur le modèles d'autres métriques (P&R), "moyenne" (pondérée : P ou R) de scores individuels

- problème de l'évaluation multi-critères
- pondération de l'un des critères / normalisation de l'influence de chaque critère

$\Rightarrow$ priorité à la bonne formation syntaxique, à la fidélité par rapport au texte source, ...

Métriques composites



BLEU-BLANC-ROUGE

Problèmes liés aux métriques

Les métriques proposées ne se comportent pas de façon homogène : "language bias problem"

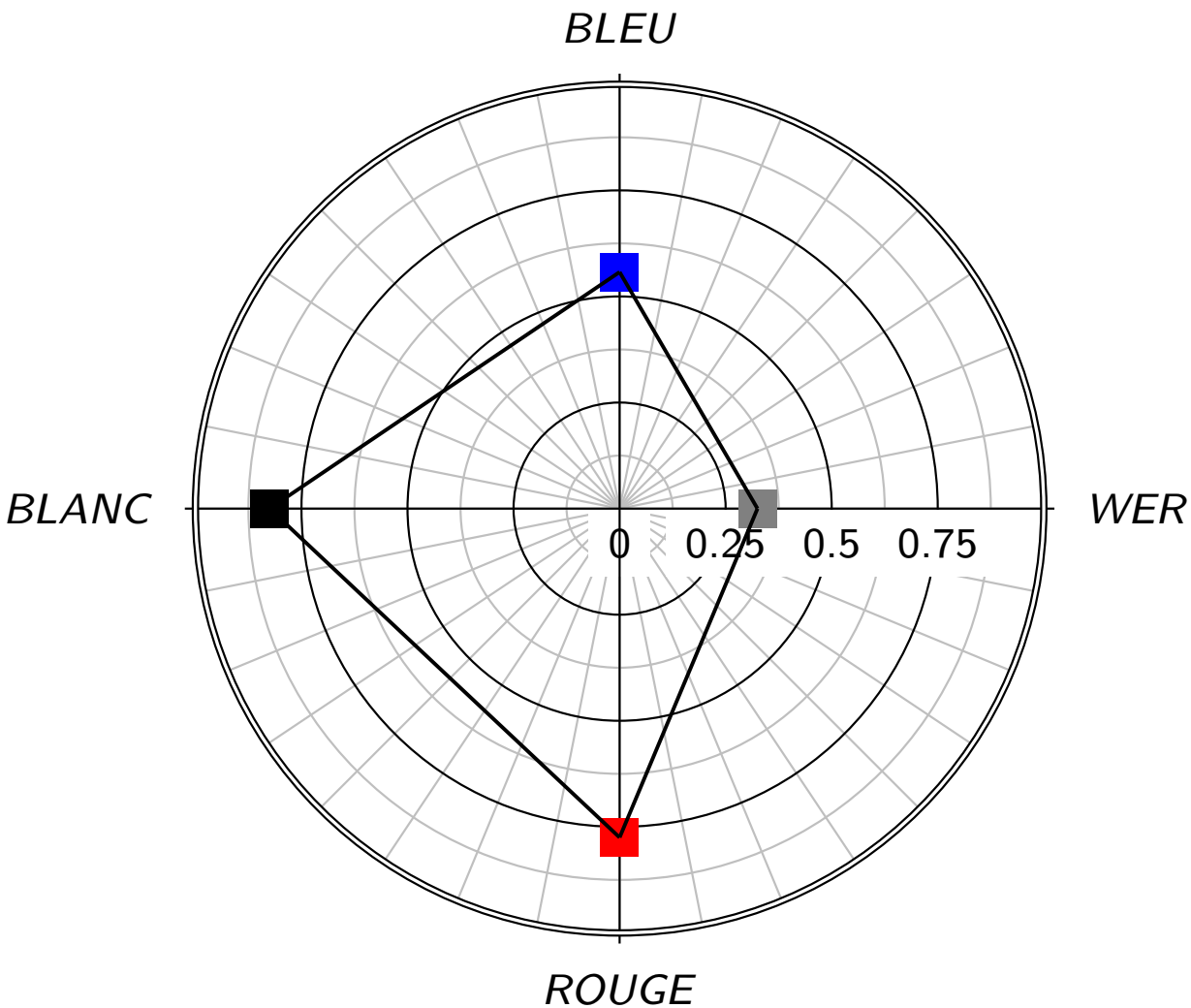
- SMT > EBMT dans le sens fr. → en., mais EBMT > SMT dans l'autre sens

Difficulté d'intégrer les métriques et de visualiser des scores composites

- métaphores graphiques à trouver, ex., "radar", toile d'araignée, etc.

une métrique par configuration de traduction ! ! !

Métriques composites ?



Évaluation par comparaison TA/BT

Les métriques de type BLEU sont perçues comme insuffisantes

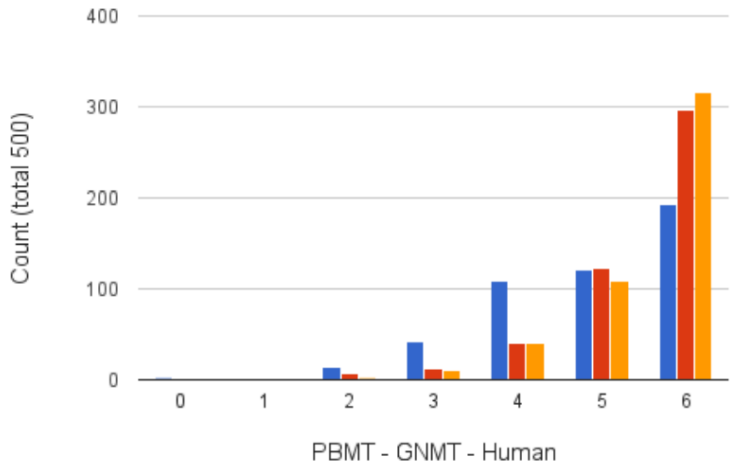


Figure 4 – Évaluation de 3 types de systèmes (500 articles wikipedia, scores de 0 à 6)

[Wu et al., 2016]

→ Traduire est une tâche complexe

→ Évaluer la qualité de la traduction est une tâche complexe

Synthèses et perspectives

Si “Que Choisir” évaluait la TA...

Limites linguistiques

Bar-Hillel souligne l'impossibilité fondamentale du projet de TA, au-delà des expériences et démonstrations particulières

- "the box is in the pen" / "the pen is in the box"
- "je prends un avocat en entrée" / "je prends un avocat pour ce litige"

$\Rightarrow$ Traitement de la polysémie : PEN_1 vs. PEN_2, $AVOCAT_1$ vs. $AVOCAT_2$
$\Rightarrow$ La question du savoir encyclopédique : en général un enclos est plus grand qu'une boîte, les humains ne mangent pas d'autres humains

Limites linguistiques

Problèmes connus de traduction

- unités polylexicales
  - collocations, expressions figées, mots composés, patrons phrastiques, etc. : “to bring the house down”, “to spill the beans”, “en avoir après qq1”, “casser sa pipe”, etc.
- “trous” lexicaux et différences de Valeur : “nuts” / “fruits à coque ?”, “leña / madera” / “bois de chauffage / d’œuvre ?”, “bouquet” / “bunch of flowers”
- faux amis : “à la mode” (fr., UK en., US en.), “have a bash” (UK/US en.), “chips” (fr. / UK en. / US en.)

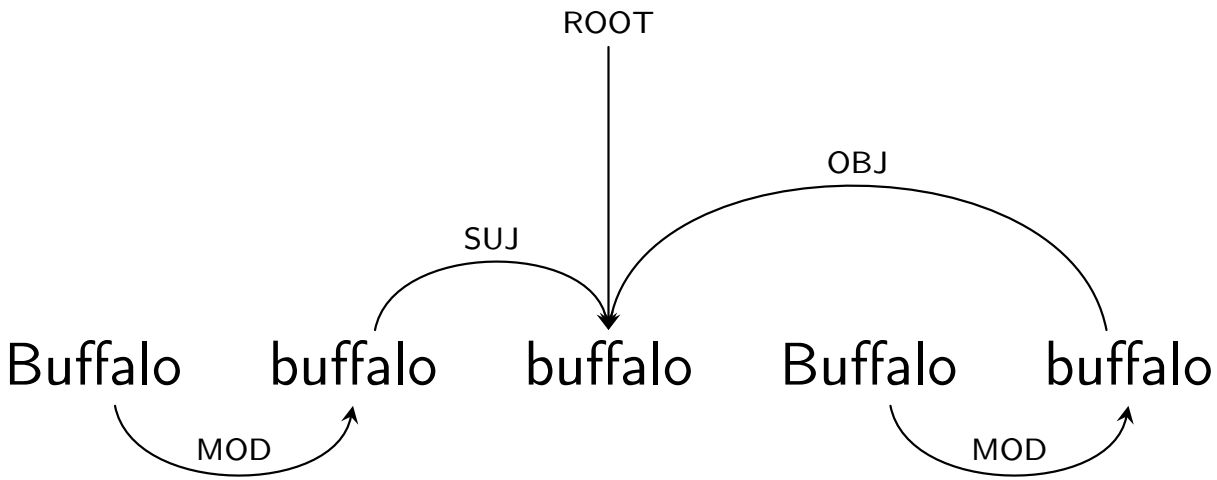
Limites linguistiques

Prise en charge de la structure syntaxique

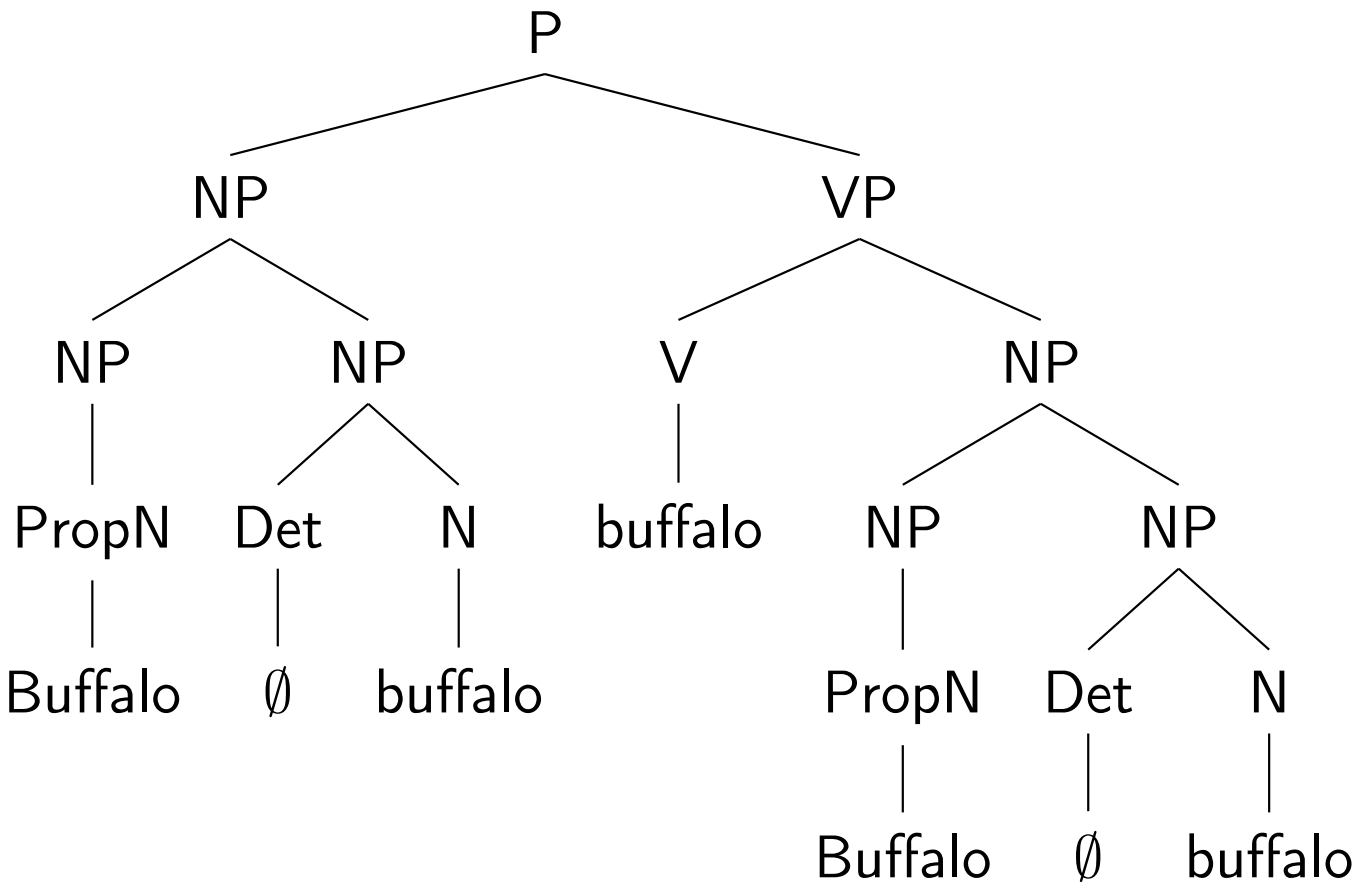
- "Buffalo buffalo buffalo Buffalo buffalo" inanalysable par la plupart des systèmes de TA
  - Google translate : idem
  - "Buffle de Buffalo de buffle de buffle de Buffalo" (Systran)

$\Rightarrow$ pas d'analyse de la structure syntaxique au-delà du syntagme (Systran) ou du segment (n-gramme, "phrase")

Limites linguistiques



Limites linguistiques



Google translate

- “the pen is in the box” → “le stylo est dans la boîte”
- “the box is in the pen” → “la boîte est|se trouve dans le stylo”

Google translate

- “je prends un avocat en entrée” / “I take an attorney(lawyer)”
- “Je prends un avocat en hors-d’œuvre” / “I take an avocado out of the hors-d’oeuvre”
- “Je prends un avocat comme hors-d’œuvre” / “I take a lawyer as an appetizer”

Google / Systran

Google

- “Le député prend la parole” / “The hon.” (update : “The member speaks”)
  - suggestion “The hon. / MP speaks”
- “The deputy takes the floor” / “Le député prend la parole”

Systran

- “Le député prend la parole” / “The deputy takes the floor”
- “The deputy takes the floor” / “Le député prend la parole”

Limites linguistiques

Défis

- unités polylexicales, expressions idiomatiques
  - "The spirit is willing but the flesh is weak" / "L'esprit est prêt, mais la chair est faible" (Google) / "L'esprit est prompt, mais la chair est faible" (Systran)
  - "Qui trop embrasse mal étreint" / "Who graps all, looses" (Google ✓✗) / "Who too embraces badly hugged" (Systran ✗) / "all covet, all lose" (Systran ✓)
    $\rightarrow$ "grasp all, lose all"
- mots inconnus
  - mots techniques, rares, registre familier/argot
  - néologismes
  - flexions rares ou nouvelles (réformes orthographiques)
  - emprunts

Limites linguistiques

Effet des réformes orthographiques :

- “L’évènement a suscité de nombreuses réactions.” / “The event provoked many reactions.” (Google) / “The event caused many reactions” (Systran)
- “L’événement a suscité de nombreuses réactions.” / “There were many reactions to the event. (Google) / “The event caused many reactions” (Systran)

Limites linguistiques

Effet des réformes orthographiques :

- “la grenouille s’installe sur son nénufar” / “the frog settles on his nenufar” (Google) / “the frog is installed on its nénufar” (Systran)

Limites linguistiques

Emprunts

- “j’ai trouvé un super jogging en soldes” / “I found a great jog on sale” ✗(Google) / “I found a super jogging in balances” ✗(Systran)

Limites linguistiques

Google et le temps qui passe :

- “J’ai (30|40...999) ans” / “I am (30|40...999) years old”
- “J’ai 1000 ans” / “I **have** a thousand years”
- “J’ai 2000 ans” / “I am 2000 years old”

Synthèse et perspectives

Ambiguity, moreover, attaches primarily to nouns, verbs, and adjectives; and actually (at least so I suppose) to relatively few nouns, verbs, and adjectives. Here again is a good subject for study concerning the statistical semantic character of languages. But one can imagine using a value of N that varies from word to word, is zero for "he," "the," etc., and which needs to be large only rather occasionally. Or would it determine unique meaning in a satisfactory fraction of cases, to examine not the 2N adjacent words, but perhaps the 2N adjacent nouns? What choice of adjacent words maximizes the probability of correct choice of meaning, and at the same time leads to a small value of N?

(Weaver, 1949)

Synthèse et perspectives

W. Weaver esquisse dans [Weaver, 1955] les pistes de recherche qui sous-tendent la TA jusqu'à nos jours :

- approches statistiques (référence aux travaux de C. Shannon)
- nécessité d'identifier les "invariants linguistiques" (Interlingua)
  - " Perhaps the way is to descend, from each language, down to the common base of human communication – the real but as yet undiscovered universal language – and then re-emerge by whatever particular route is convenient."
  - conviction que le projet est réalisable dans un cadre formalisé : "Such a program involves a presumably tremendous amount of work **in the logical structure of languages** before one would be ready for any mechanization."

Synthèse et perspectives

Les annonces faites par Google laissent espérer (craindre ?) un saut qualitatif

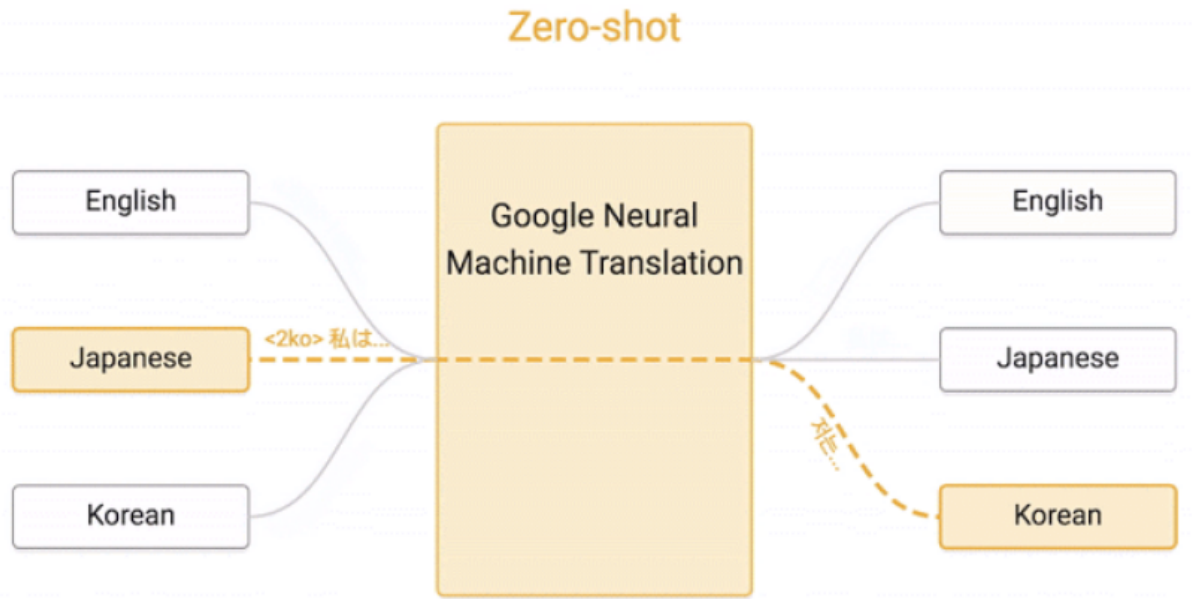
- "Neural MT" (NMT)
  - intégration de données textuelles et d'images au cours du processus de paramétrage
  - dépassement du goulot d'étranglement dû à l'explosion combinatoire liée à l'approche par transfert
    - "Zero-Shot Translation" : traduction associative (= transitivité)
    - généralisation des modèles de traduction au-delà des paires de langues fournies en apprentissage
    - "low resource language improvement" : équipement à peu de frais des langues sous-dotées

Google "zero-shot" translation

The success of the zero-shot translation raises another important question : Is the system learning a common representation in which sentences with the same meaning are represented in similar ways regardless of language — i.e. an "interlingua" ?

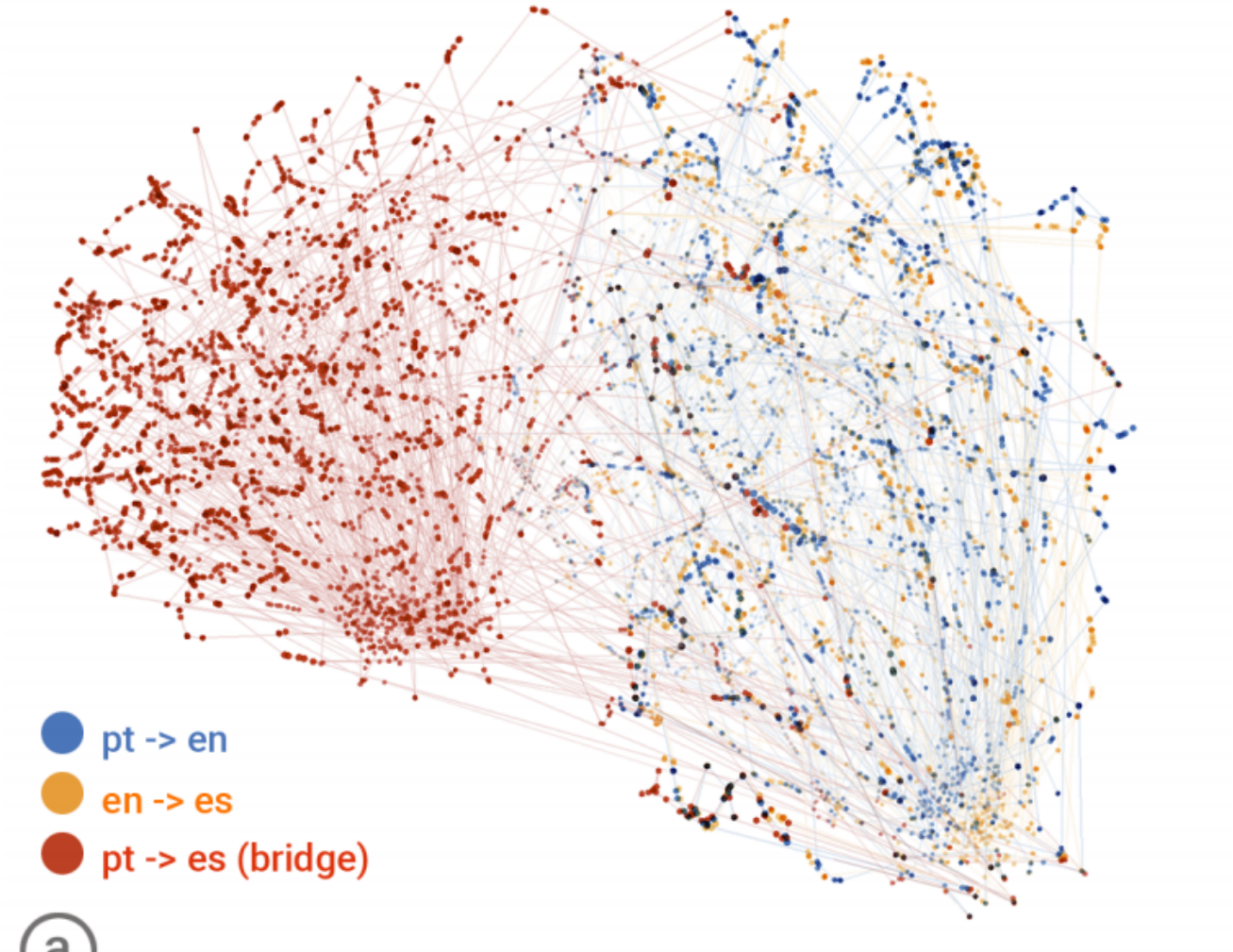
[Zero-Shot Translation with Google's Multilingual Neural Machine Translation System]

Google "zero-shot" translation

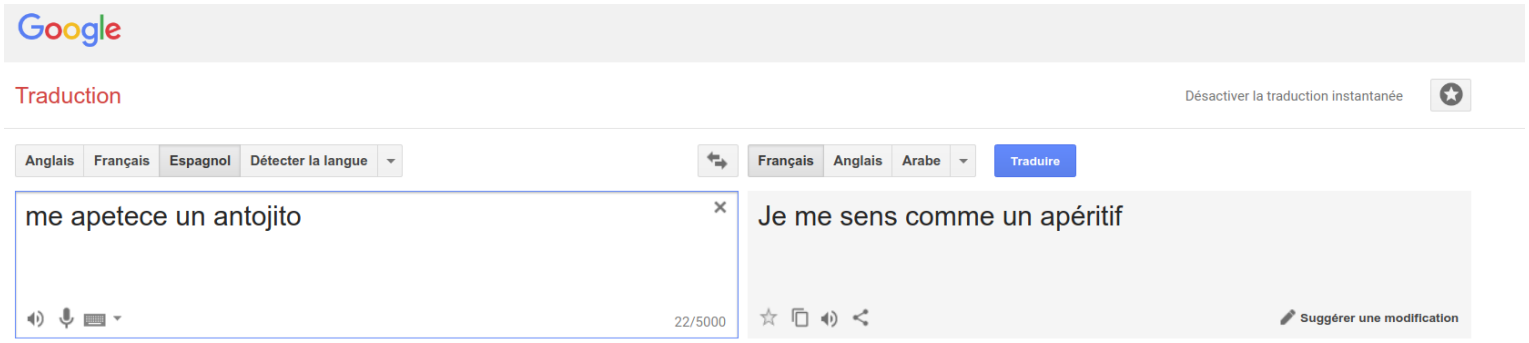


Google "zero-shot" translation

"Evidence for an Interlingua" [Johnson et al., 2016]



Google : “zero-shot” translation



“I feel like” $\rightleftharpoons$ “me apetece” $\rightarrow$ “je me sens comme” ?


36. PLATO LAS DELICIAS (CHICHARRON DE CARNE, CHIMICHURRI, TORTILLA
PLATE THE DELIGHTS (CHICHARRON OF MEAT, CHIMICHURRI, HOME
37. FAJITAS DE POLLO, (CHILE DULCE, CEBOLLA, PAPAS FRIT
STRIPS OF CHICKEN (SWEET CHILI, ONION, French fries AN
38. PESCADO ENTERO (PATACONES Y ENSALADA ORDEN D
FISHED ENTIRELY (PATACONES AND SALAD ORDER OF R
39. LENGUA EN SALSA (ARROZ BLANCO ENSALADA SOLO
LANGUAGE IN SAUCE (RICE WHITE SALAD SINGLE

Références bibliographiques I


Banerjee, S. and Lavie, A. (2005).
Meteor : An automatic metric for mt evaluation with improved correlation with human judgments.
In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.


Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., and Roossin, P. (1988).
A statistical approach to language translation.
In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 71–76. Association for Computational Linguistics.

Références bibliographiques II

Buchanan, B. G. and Shortliffe, E. H., editors (1984).
Rule-Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project.
Addison-Wesley Series in Artificial Intelligence. Addison Wesley.

Doddington, G. (2002).
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.
Morgan Kaufmann Publishers Inc.

Hebb, D. O. (1949).
The organization of behavior : A neuropsychological theory.
Wiley & Sons, New York.

Références bibliographiques III

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2016).
Google's multilingual neural machine translation system : Enabling zero-shot translation.
CoRR, abs/1611.04558.

Koehn, P. (2009).
Statistical machine translation.
Cambridge University Press.

Lin, C.-Y. (2004).
Rouge : A package for automatic evaluation of summaries.
In *Text summarization branches out : Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Références bibliographiques IV

Lin, C.-Y. and Och, F. J. (2004).
Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics.
In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

McCulloch, W. S. and Pitts, W. (1943).
A logical calculus of the ideas immanent in nervous activity.
The bulletin of mathematical biophysics, 5(4) :115–133.

Nagao, M. (1984).
A framework of a mechanical translation between japanese and english by analogy principle.
Artificial and human intelligence, pages 351–354.

Références bibliographiques V


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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.


Pierce, J. R. and Carroll, J. B. (1966).
Language and machines : Computers in translation and linguistics.
National Academy of Sciences/National Research Council.

Références bibliographiques VI


Rosenblatt, F. (1958).
The perceptron : a probabilistic model for information storage and organization in the brain.
Psychological review, 65(6) :386.


Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.
CoRR, abs/1409.3215.


Weaver, W. (1955).
Translation (1949) in machine translation of languages.

Références bibliographiques VII

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation.
CoRR, abs/1609.08144.

Webographie

- http ://ohmy-box.info/linguistics/fr/evaluations_metriques.php
- http ://ohmybox.info/linguistics/fr/evaluations_humaines.php
- http ://asiya.lsi.upc.edu/demo/asiya_online.php
- http ://www.mt-archive.info
- http ://www.undl.org/
- http ://people.dbmi.columbia.edu/ ehs7001/Buchanan-Shortliffe-1984/MYCIN%20Book.htm
- http ://blog.teamleadnet.com/2017/08/building-statistical-machine-translation.html