NRC·CNRC

Institute for Information Technology

Entraînement, usagers, évaluation, performance, et avenir des systèmes TAS

Roland Kuhn

Septembre 30, 2010

National Research Council Canada
Conseil national de recherches Canada
Canada

NRC·CNRC
Institute for Information Technology

Rappel: structure détaillée d'un système TAS

Corpus bilingue aligné

S  *T* 

Corpus unilingue (langue cible)

New York Times

S: *Mais où sont les neiges d'antan?*

✂ Segmenter

mais? mais où?
mais où sont?
où? où sont? …

Chercher →

Modèle de traduction de segments

Table de segments

P(but | mais)=0.7
P(however | mais)=0.3
……
P(where are | où sont) =1.0
……

Coller & scorer ↓

Sortie du système TAS

T1: *but where are the snows of yesteryear?* **P = 0.41**
T2: *however, where are yesterday's snows?* **P = 0.33**
T3: *hey - where did the old snow go?* **P = 0.18**
…

Modèle de langue *n*-gramme

aalborg is a	→	3
…		
many people say	→	4,548
…		
where are the	→	931
…		
zulu is the	→	4

Scorer ↓

Autres info.

Longueur (1 mot anglais ≈ 1.2 mots français)
Réordonnancement
Etc.

Sco rer ↓

Décodeur

1. Couper ***S*** en segments: ✂
2. Chercher les segments dans la table
3. Coller ensemble les segments de langue cible → hypothèses
4. Scorer les hypothèses

NRC·CNRC
Institute for Information Technology

Règles pratiques pour l'entraînement d'un système TAS par segments

- Vous aurez besoin d'un grand corpus bilingue aligné pour entraîner le modèle de traduction de segments (MT) et d'un grand corpus unilingue en langue cible pour entraîner le modèle de langue (ML).

- Ces deux corpus doivent :
  1. être dans le même domaine que les textes qu'on aimerait traduire avec le système. Par ex., si on veut traduire des documents pour le Ministère de l'agriculture, on devrait entraîner le système sur des traductions qui viennent de ce ministère (entraîner sur des traductions d'un autre ministère risque de donner de très mauvais résultats).
  2. avoir une taille suffisante – le minimum est probablement 1 million de mots de langue cible pour le MT et le ML (sauf pour des domaines avec un vocabulaire très spécialisé, tels la météo).
  3. être de bonne qualité (pas trop de fautes d'orthographe, *etc.*) – mais en général, **la quantité est plus importante que la qualité**. Mieux vaut 100 millions de mots de textes de qualité médiocre que 1 millions de mots de haute qualité.

- Une mémoire de traduction est un bon départ: on entraîne le MT sur les paires de phrases, et on entraîne le ML sur les phrases en langue cible.

NRC·CNRC
Institute for Information Technology

Règles pratiques pour l'entraînement d'un système TAS par segments

- Situation typique: on a un MT et un ML entraînés sur des corpus qui viennent du domaine (par ex., qui viennent d'une mémoire de traduction spécialisée au Ministère de l'agriculture) mais on a aussi de données bilingues et unilingues qui viennent d'autres domaines. Est-ce qu'on peut exploiter ces autres corpus?
- OUI. Le système peut se servir de plusieurs MT et ML, mettant un poids sur chacun. Une procédure « **MERT** » (« Minimum Error Rate Training ») apprend automatiquement le poids à donner à chaque MT et ML, si on lui donne un échantillon de paires de phrases qui viennent du domaine (« **Dév** » = corpus de développement). Normalement, MERT va donner les plus grands poids au MT et ML du domaine.

Mém. trad. – Min. de l'agriculture

Mém. trad. - Min. de finance

La Presse

MT_{ag} ML_{ag} MT_{fi} ML_{fi} ML_{LP}

Dév (dom = agriculture)

The chickens are healthy
Les poulets sont en bonne santé.
…

→ **MERT** → **Poids**

$$\underline{0.7}^{*}\, MT_{ag} \;\&\; \underline{0.3}^{*}\, MT_{fi} \;\&\; \underline{0.6}^{*}\, ML_{ag} \;\&\; \underline{0.1}^{*}\, ML_{fi} \;\&\; \underline{0.3}^{*}\, ML_{LP}$$

NRC·CNRC
Institute for Information Technology

Considérations computationnelles

- L'entraînement d'un système TAS demande beaucoup plus de calcul que l'utilisation du système pour la traduction. Rappelons que l'entraînement a trois phases:
  1. l'entraînement du modèle de langue (ML) – ne demande pas beaucoup de calcul (à moins que le corpus monolingue cible ne soit pas grand).
  2. l'entraînement du modèle de traduction de segments (MT) – demande beaucoup de calcul.
  3. MERT – demande beaucoup de calcul.

- Conséquence pratique: pour entraîner un bon système TAS, on a besoin de beaucoup de machines. Une fois le système entraîné, une seule machine peut suffire pour faire tourner le système.

  ⇨ Une entité/compagnie avec beaucoup de machines et avec accès aux données d'entraînement peut s'occuper de l'entraînement;
  le système entraîné est livré à un client (par ex., une compagnie de traduction) qui s'occupe de faire des traductions et qui n'a donc pas besoin d'acheter beaucoup de machines puissantes.

  * Nous (le groupe PORTAGE du CNRC) faisons exactement ça – nous avons accès à une grande grappe de calcul, beaucoup plus puissante que les machines de nos « clients ».

NRC·CNRC
Institute for Information Technology

Applications et usagers de la TAS

- Le gouvernement américain aimerait surveiller des développements dans certains autres pays, mais n'a pas assez d'analystes qui maîtrisent les langues en question ⇨ beaucoup de $$$ pour recherche en TAS:
  * → **anglais** - où * = { arabe, chinois, coréen, dari, farsi, pushtou ….}.

  Ennemis potentiels des ÉU

  Pour ces applications de veille, **la fidélité est plus importante que la fluidité**.

- Les grandes sociétés américaines (par ex., IBM, Microsoft) font traduire leurs manuels: **anglais → *.** Politique de Microsoft: sortir la première version vite (avec TAS), post-édition humaine par après.

- La Communauté européenne: (CE) le plus grand « consommateur » de traductions au monde. 23 langues officielles ⇨ 506 paires de langues directionnelles (le nombre est en croissance). 5,000 traducteurs internes + 2,000 personnel de support. Coût: 1% du budget du CE = 3 milliards € par année. Techniques utilisées: mémoire de traduction + TABR + projet Exodus (anglais → portugais, anglais → français).

- Language Weaver*:* technologie de TAS sur le site du client.

- PORTAGE: notre système TAS

NRC·CNRC
Institute for Information Technology

Applications et usagers de la TAS

Dr. Franz Josef Och

Inventeur principal de la TAS par segments
Chef de l'équipe « Google Translate »
Opposant féroce des règles, de la syntaxe, *etc.*

- Google Translate:
  - TABR (Systran) avant 2007, TAS depuis
  - 57 langues (sept. 2010), y compris certaines langues « petites » (créole, islandais)
  - Gratuit.
  - Utilisateurs: le grand public. Surtout des applications de veille, mais parfois utilisé pour aider ou remplacer le processus de traduction par un être humain.
  - Souvent utilisé en catimini. Par ex., écoliers qui trichent dans leurs devoirs…… mais aussi certains traducteurs professionnels!
  - Sources de données bilingues d'entraînement (en général, plus difficiles à trouver que les données unilingues en langue cible):
    * le « Linguistic Data Consortium » (banque de données académiques, ÉU)
    * les Nations-Unies (documents dans les 6 langues officielles = anglais, arabe, chinois, français, russe, espagnol)
    * surtout, le Web (Google est bien positionné pour recueillir ces données!)
    * éventuellement, Google books?

Entrevue avec Franz-Josef Och, « L.A. Times », mars 2010:
http://latimesblogs.latimes.com/technology/2010/03/the-web-site-translategooglecom-was-done-in-2001-we-were-just--licensing-3rd-party-machine-translation-technologies-tha.html

NRC·CNRC
Institute for Information Technology

Aspects de la qualité des systèmes TAS

- Les systèmes TAS progressent vite: meilleurs algorithmes, plus de données multilingues sur le Web, machines plus puissantes.
- La qualité des systèmes est très variable: même avec des quantités comparables de données d'entraînement, certaines paires de langues (par ex., chinois-anglais, japonais-anglais) ne donnent que des sorties très médiocres. Pour ces paires de langues, seulement les applications les moins ambitieuses – par ex., la veille – sont envisageables à court terme.
- Pour certaines paires de langues, la qualité est assez bonne pour envisager des applications plus ambitieuses: pour ces paires de langues telles (par ex.) anglais-français, français-espagnol, on propose la TAS comme aide à la productivité des traducteurs humains.
- Par exemple, pour ces paires de langues, il est possible que la TAS suivie d'une post-édition humaine rende les traducteurs plus productifs – essais en cours! Bien sûr, cette procédure ne serait utile que dans les domaines qui: 1. génèrent beaucoup de données en format électronique 2. ont tendance à utiliser les mots et les séquences de mots de façon très répétitive.
- Le domaine gouvernemental est idéal: là, la langue de bois règne! Par contre, les textes littéraires ne se prêtent pas à la TAS.

NRC·CNRC
Institute for Information Technology

La qualité des systèmes TAS: fidélité et fluidité

- Pour comparer les systèmes de TA, il faut mesurer la qualité des traductions qu'ils produisent.
- Les critères traditionnels pour les traductions sont la fidélité (une traduction fidèle contient les même informations que la source) et la fluidité (une traduction fidèle donne l'impression d'avoir été écrite directement dans la langue cible).
- Pour les applications de veille, la fidélité est plus importante que la fluidité. Pour les applications plus ambitieuses telles l'aide aux traducteurs humains, les deux sont importantes.
- Quand les gens évaluent la fidélité et la fluidité d'une traduction, ces deux critères finissent par être corrélés. Même si elle contient toutes les informations de la source, une traduction qui n'est pas fluide sera jugée « pas fidèle » parce que c'est trop difficile de trouver ces informations dans la traduction.
- Beaucoup de fluidité sans fidélité est très dangereux pour le lecteur unilingue (en langue cible). Ça arrive parfois avec Google Translate! Exemple inventé:

Source: **Marie plâit à Jean.**
Trad. 1: **Mary likes to John.** ← *pas fluide; le lecteur se méfie.*
Trad. 2: **Mary likes John**. ← *fluide; le lecteur pense que c'est fiable.*

NRC·CNRC
Institute for Information Technology

Évaluation des systèmes TAS: mesures humaines

- La façon la plus évidente de comparer et d'évaluer les TA est d'utiliser le jugement des êtres humains (les juges). Ce sont des mesures humaines.
- Par ex., on pourrait demander aux juges de donner à chaque phrase produite par le système TA un score de 1, 2, 3, 4, ou 5 pour la fidélité et aussi un score 1-5 pour la fluidité. On met un texte source de ~1,000 phrases dans chaque système; ce texte s'appelle le « texte test ». Le système qui a la meilleure moyenne de fidélité et de fluidité sur le texte test gagne la compétition.
- Ou on pourrait demander aux juges d'ordonner les sorties de plusieurs systèmes TA: meilleure TA d'une phrase → pire TA (matchs nuls permis).
- Idéalement, les juges seraient des traducteurs experts sur la paire de langues en question. Mais les traducteurs coûtent cher (et ils détestent ces tâches). Très souvent, on fait faire par des traducteurs des traductions de référence du texte test. Les juges seront des gens – très souvent unilingues - qui connaissent la langue cible mais pas nécessairement la langue source. Ils jugeront chaque sortie d'un système TA à partir des références, pas à partir de la source. Les juges unilingues coûtent moins cher …

NRC·CNRC
Institute for Information Technology

Évaluation des systèmes TAS: mesures humaines

- Problème 1: dans les évaluations humaines où chaque évaluateur donne un score pour la fidélité et un score pour la fluidité, il y a toujours beaucoup de divergences entre les scores d'évaluateurs différents. Mon jugement de ces deux qualités est probablement différent du tien. Même problème pour les évaluations où on ordonne les sorties TA.
- Problème 2: ces approches sont trop éloignées de la tâche concrète pour laquelle on aimerait utiliser un système de TA. Donc, il y a d'autres mesures humaines qui sont plus ciblées sur une tâche concrète.

Évaluation des systèmes TAS: mesures humaines

- Mesure ciblée sur les applications de veille: **DLPT*** (basée sur le **D**efense **L**anguage **P**roficiency **T**est du militaire américain). On demande aux évaluateurs de lire la TA d'un groupe d'articles et de subir un test à choix multiples sur le contenu de chaque article. Si les évaluateurs qui ont lu les traductions de **Système 1** obtiennent un score plus élevé que les évaluateurs humaines qui ont lu les traductions de **Système 2**, on conclut que **Système 1** est un meilleur outil pour les analystes militaires. DLPT* favorise la fidélité.

* Voir http://www.ll.mit.edu/publications/journal/pdf/vol18_no1/18_1_2_Jones.pdf

Texte source

مصر للطيران قد تعاود غدا الاربعاء رحلاتها الى ليبيا
6 القاهرة - 4 اعلن - (اف ب)
مسؤول في شركة الشطوط المصرية "مصر للطيران اليوم الثلاثاء ان شركة "للطيران قد تضيف اعتبارا من يوم غد الاربعاء رحلاتها الى ليبيا اثر قرار مجلس الامن الدولي تعليق الحظر المفروض على ليبيا

Formulation de l'examen

Examen

Q.1: L'article discute les événements dans quels pays?
A. Le Maroc
B. La Tunisie
…

Système TA

Traduction

Adobe, the Adobe logo, Acrobat, the Acrobat logo, Acrobat Capture, Adobe Garamond, Adobe Intelligent Document Platform, Adobe PDF, Adobe Reader, Adobe Solutions Network, Alfin, Distiller, ePaper, Extreme, FrameMaker, Illustrator, InDesign, Minion, Myriad, PageMaker, Photoshop, Poetica, PostScript, and XMP are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Microsoft and Windows are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Apple, Mac, Macintosh, and Power Macintosh are trademarks of Apple Computer, Inc., registered in the United States and other countries. IBM is a registered trademark of IBM Corporation in the United States. Sun is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries. UNIX is a registered trademark of The Open Group. XVG is a trademark of the World Wide Web Consortium; marks of the W3C are registered and held by its host institutions MIT, INRIA and Keio. Helvetica and Times are registered trademarks of Linotype-Hell AG and/or its subsidiaries. Arial and Times New Roman are trademarks of The Monotype Corporation registered in the U.S. Patent and Trademark Office and may be registered in certain other jurisdictions. ITC Zapf Dingbats is a registered trademark of International Typeface Corporation. Ryumin Light is a trademark of Morisawa & Co., Ltd. All other trademarks are the property of their respective owners.

Réponses

Q. 1: B
Q. 2: D
…

NRC·CNRC Institute for Information Technology

Évaluation des systèmes TAS: mesures humaines

- Mesure ciblée sur l'aide aux traducteurs humains: **HTER** = **H**uman-mediated **T**ranslation **E**dit **R**ate. Après avoir lu les références/la référence, l'évaluateur corrige la sortie du système pour qu'elle soit fidèle et fluide (pas qu'elle soit identique aux références!) Avantage: reflète l'effort de post-édition; favorise surtout la fluidité.
- Définition:
  score HTER = (# de changements dans la TA)/(# de mots dans la référence).

Source

Mon ami se comporte comme un idiot.

Système TA

My buddy himself behaving as idiot.

Référence

My friend is acting like an idiot. **(7 mots)**

Corrections

My buddy himself behaving as idiot.

is **like an**

Version corrigée

My buddy is behaving like an idiot.

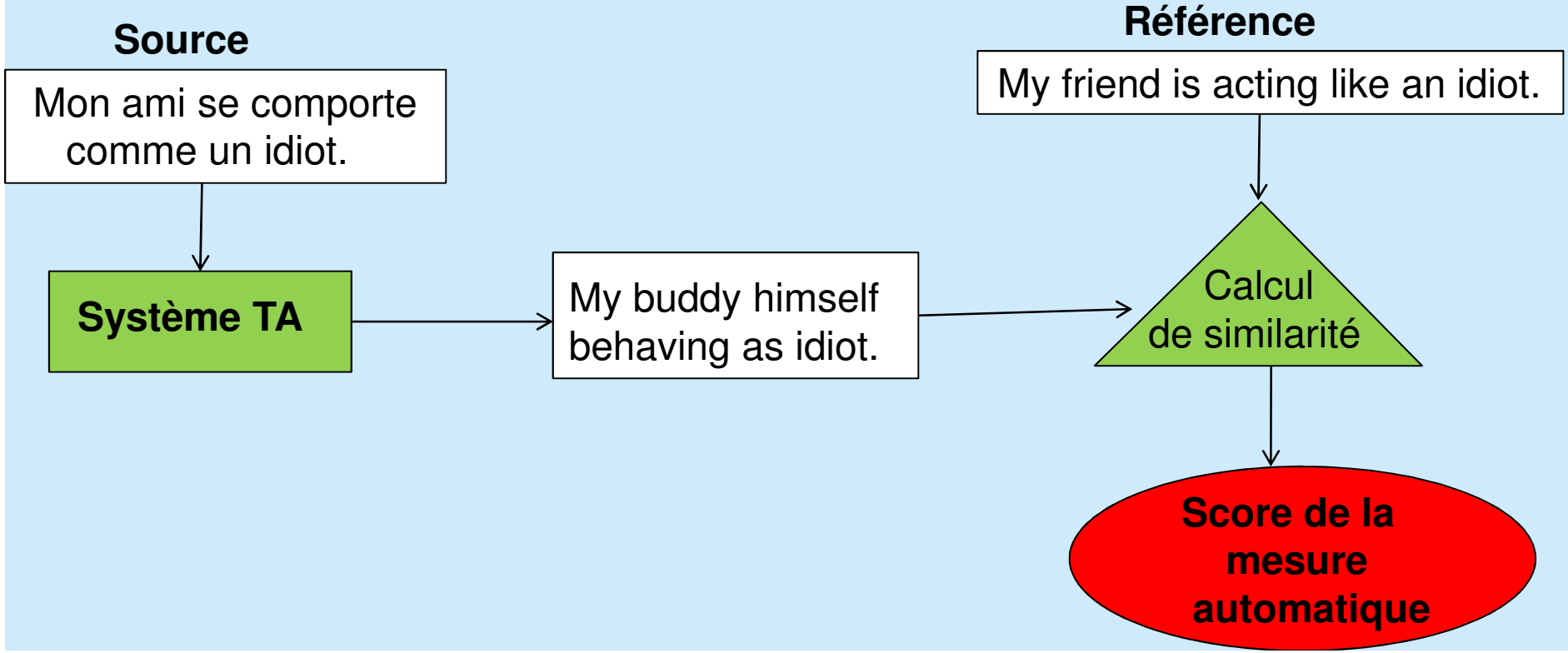
Score HTER

5 changements/7 mots = **0.71**

NRC·CNRC
Institute for Information Technology

Évaluation des systèmes TAS: mesures automatiques

- Problème avec les mesures humaines: souvent, on a besoin de **beaucoup** d'évaluations – les mesures humaines prennent trop de temps et coûtent trop cher.
- Exemple classique: MERT = algorithme de calcul des poids optimaux. Cet algorithme demande la comparaison des milliers de petites variations du même système – impossible si on utilise des évaluateurs humains.
- Les mesures automatiques font tous un calcul de similarité entre la sortie d'un système TAS et une ou plusieurs références.



NRC·CNRC
Institute for Information Technology

Une liste partielle des mesures automatiques

- WER = « word error rate »: emprunté aux recherches sur la reconnaissance de parole. Distance d'édition entre la TA et la référence; pénalise les variations d'ordre légitimes.

- TER = « translation edit rate »: Distance d'édition avec mécanisme facilitant le déplacement de blocs de mots. Intuitivement: coût de transformer la TA en traduction de référence.

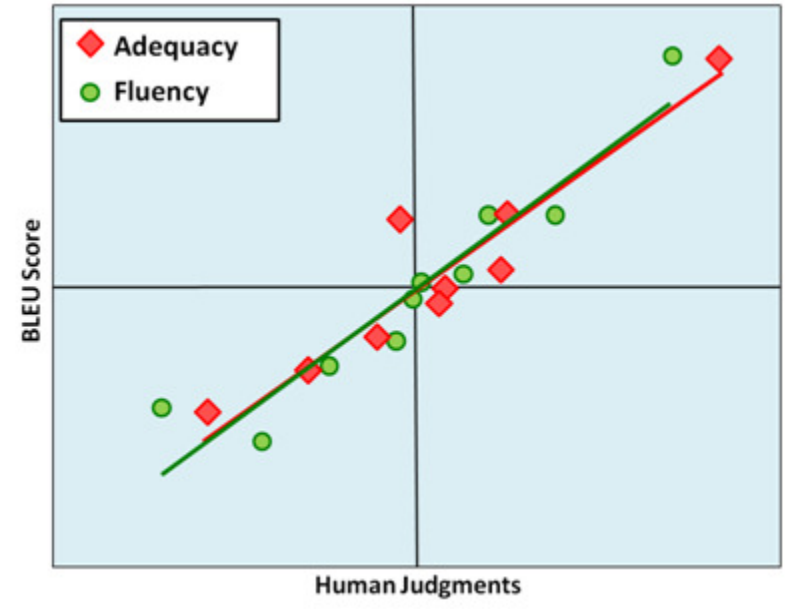
  NOTE: WER et TER ressemblent beaucoup à HTER, mais il n'y a pas d'évaluateur humain.

- **BLEU**: de loin la mesure la plus utilisée. Moyenne de similarité en *n*-grammes pour différentes valeurs de n (typiquement, n = [1-4]).
  Unigrammes → ~ fidélité ; *n*-grammes → ~ fluidité.
  Pénalité de brièveté sur les traductions avec longueur trop différente de celle de la référence.

NRC·CNRC
Institute for Information Technology

Avantages des mesures automatiques

- Rapides et peu coûteuses
- Pour le développement: comparaison automatique de multiples variantes du système
- Idéales pour MERT
- Produisent un ranking de systèmes souvent fortement corrélé avec leur ranking humain (le ranking de phrases individuelles est beaucoup moins correlé!)



NRC·CNRC
Institute for Information Technology

Désavantages des mesures automatiques

- La corrélation avec les jugement humains est très imparfaite pour les phrases individuelles.
  Réf = The man spoke rudely to me.
  TA 1 = The man spoke politely to me. *(Presque tous les mots sont identiques).*
  TA 2 = He was insolent. *(0 mots identiques).*
  → **TA 1 obtient un bien meilleur score BLEU!**
- La corrélation n'est suffisamment bonne que sur des textes entiers (et non des phrases)
- La performance relative des différentes métriques tend à varier avec les conditions de d'évaluation (couple de langues, genre de texte, nombre de traductions de référence ,*etc.).*

NOTE: la longueur du texte test et le nombre de références sont importants– par ex., un score **BLEU** calculé sur un texte de 5,000 phrases avec 4 références chacune est beaucoup plus fiable qu'un score **BLEU** calculé sur 200 phrases avec 1 référence chacune.

NRC·CNRC
Institute for Information Technology

METEOR: une mesure automatique plus intelligente

- **METEOR** utilise la lemmatisation (conversion d'un mot → sa racine) et un dictionnaire de synonymes pour faire un calcul de similarité plus intelligent.

  Réf = I walked quickly down the street.

  TA = My walk along the road was quick.

  → **Score BLEU très bas (les deux phrases n'ont que le mot « the » en commun)**

  → **METEOR: lemma(« walked ») = « walk », lemma(« quickly ») = « quick », « street » & « road » sont des synonymes ⇨**

  Score METEOR plus favorable à cette traduction que BLEU.

- Désavantage de **METEOR**: ce score est beaucoup trop lent à calculer pour utilisation dans MERT.

NRC·CNRC
Institute for Information Technology

Évaluations internationales

- Il y a plusieurs évaluations internationales régulières pour juger la qualité des systèmes TA. N'importe quel groupe de recherche ou compagnie peut participer à ces concours.

- Évaluation NIST: américaine; a lieu presque chaque année. Langues: (surtout) arabe → anglais, chinois → anglais; (un peu de) farsi → anglais, urdu → anglais. Critère principal: automatique (BLEU).

- Évaluation WMT: européenne; a lieu chaque année. Langues: anglais → {allemand, français, espagnol, tchèque}, {allemand, français, espagnol, tchèque} → anglais. Critère principal: (humain) chaque évaluateur ordonne les TA de 1 (meilleur) → 5 (pire).

- Conclusions:
  - certaines paires de langues sont beaucoup plus difficiles
  - les mesures automatiques sont injustes envers les systèmes TABR
  - TABR obtient de meilleurs scores humains que TAS dans certains cas (par ex., quand l'allemand est la langue source ou la langue cible)
  - en général, les meilleurs systèmes actuels sont des systèmes TAS

NRC·CNRC Institute for Information Technology

La qualité de la TAS dépend de la paire de langues

- On a tendance à parler comme si la TAS obtient les résultats semblables sur n'importe quelle paire de langues. Même si on entraîne tous les systèmes sur la même quantité de données, c'est loin d'être vrai!
- Par ex, **anglais ↔ français** est une paire qui sont relativement faciles pour les systèmes TAS. Les résultats ne sont pas parfaits, mais on peut au moins envisager la TAS comme aide aux traducteurs dans ce cas.
- Les meilleurs systèmes TAS du monde n'arrivent pas à obtenir des résultats acceptables pour **chinois ↔ anglais** ou **japonais ↔ anglais**. Pour ces paires, on est loin de pouvoir bâtir quelque chose d'utile pour les traducteurs. PORTAGE était 3ième (sur 19 systèmes) dans la dernière évaluation NIST du **chinois → anglais**. Exemple:

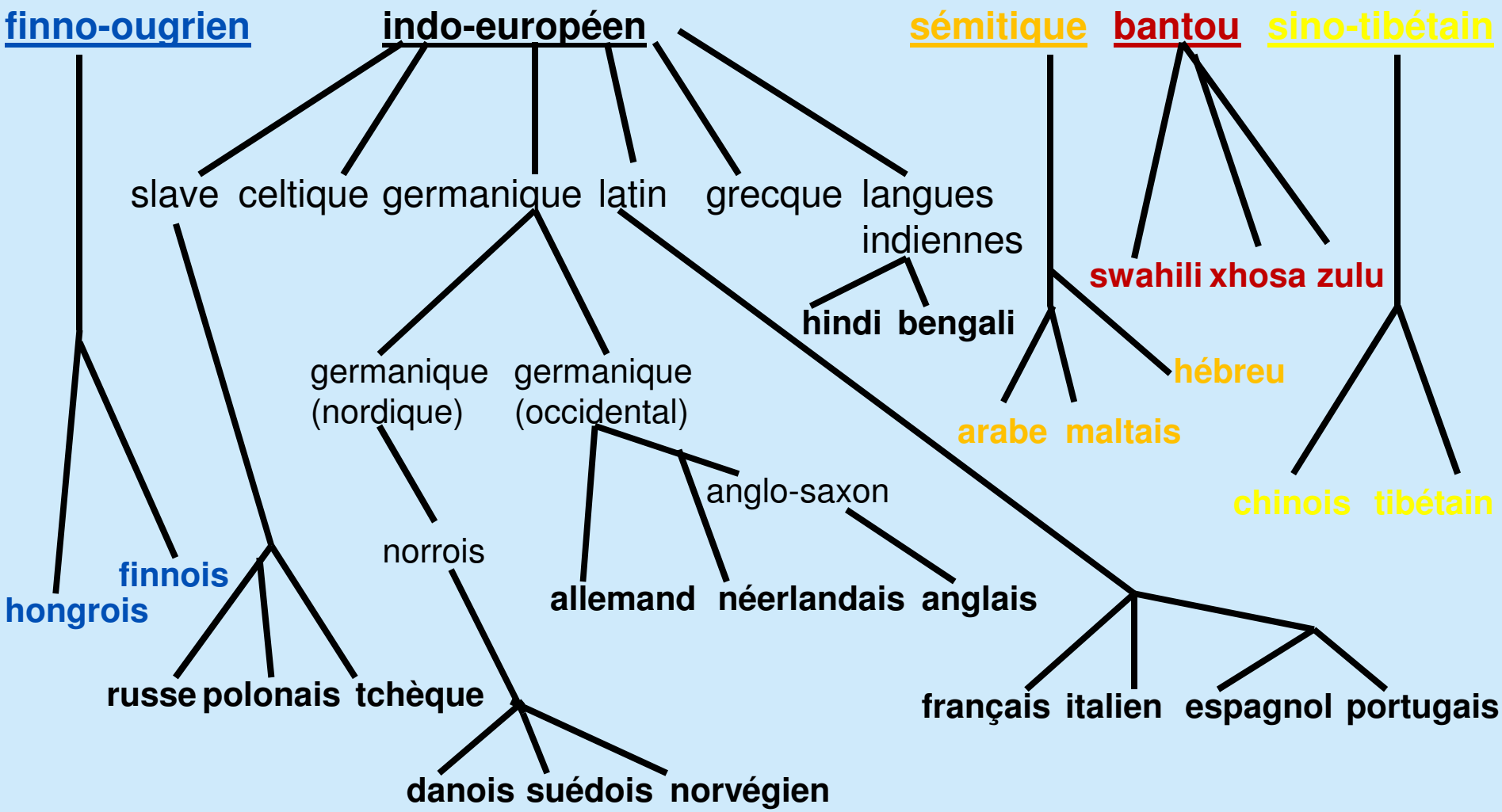
Source: 霍 夫 曼 曾 沈 迷 药物 幸 及时 醒悟 开创 演 艺 事业
Réf: Having Succumbed to Drugs, Fortunately Hoffman Sobered in Time to Start Acting Career

↓

PORTAGE (TAS): Hoffman previously enamored drug fortunate to create a timely wake up performing career

NRC·CNRC
Institute for Information Technology

Hypothèse: la qualité de la TAS dépend de la distance généalogique entre les langues



NRC·CNRC
Institute for Information Technology

Remarques sur cet arbre généalogique des langues

- Les évidences utilisées pour construire ce genre d'arbre sont:
  1. des documents anciens, inscriptions, *etc*. (par ex., on a beaucoup de sources écrites pour retracer l'histoire du français et l'anglais)
  2. la syntaxe, qui garde presque toujours des traces de l'origine d'une langue.
  3. les mots pour les concepts de base - par ex., pour « mère », « père », « bras », « œil », « eau », « pain » - ont eux aussi une forte tendance à être conservés.

- La géographie n'est pas un bon guide. Le suédois et le finnois – langues avoisinantes, dans le nord de l'Europe - n'ont rien en commun (le finnois n'est pas une langue indo-européenne). Le suédois a plus en commun avec le bengali, langue parlée dans l'est de l'Inde.

- Il reste plusieurs langues qui sont solitaires dans leur famille, ou dont la famille n'est pas connue. Exemples: 1. le basque – langue de l'Espagne et de la France – est une langue solitaire (ni indo-européen, ni finno-ougrien, ni sémitique).
  2. l'origine du japonais est inconnue (certains linguistes pensent que c'est un « cousin » du coréen, mais tous les linguistes coréens et japonais détestent l'idée).

- Prédictions pour la TAS: chinois → anglais = **lamentable**;
  arabe → anglais = **lamentable**; anglais ↔ allemand = **très bon**;
  français ↔ {italien, espagnol, portugais} = **très bon**;
  anglais ↔ français = **médiocre**.

NRC·CNRC Institute for Information Technology

Prédictions vs. réalité

Prédictions:

chinois → anglais = **lamentable**; **VRAI**
arabe → anglais = **lamentable**; **FAUX** (c'est bon et on ne sait pas pourquoi)
anglais ↔ allemand = **très bon**; **FAUX** (c'est médiocre)
français ↔ {italien, espagnol, portugais} = **très bon**; **VRAI**
anglais ↔ français = **médiocre**; **FAUX** (c'est bon)

Expériences de Koehn sur 462 paires (données identiques) – scores BLEU

Cible → Source↓	allemand	anglais	espagnol	finnois	français	maltais
allemand	X	**53.6**	47.1	29.5	39.4	19.8
anglais	46.8	X	**55.2**	38.6	**50.1**	39.8
espagnol	42.7	**60.0**	X	28.5	**51.3**	24.6
finnois	36.0	49.3	39.7	X	29.5	19.4
français	45.1	**64.0**	**60.9**	30.0	X	25.3
maltais	37.2	**72.1**???	48.7	25.8	42.4	X

NRC·CNRC
Institute for Information Technology

Prédictions vs. réalité

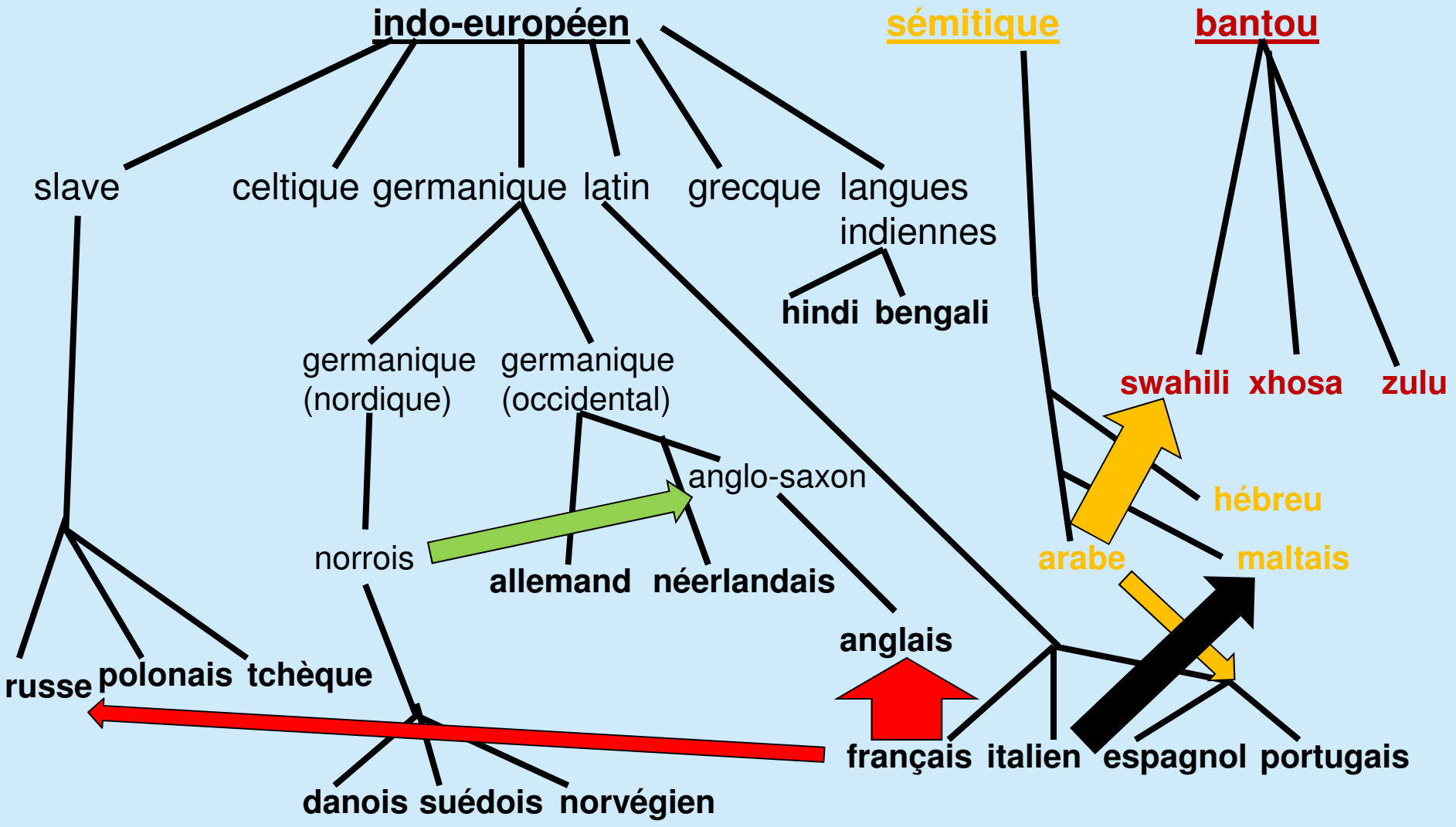
- Les 8 scores les plus élevés:
  maltais → anglais = **72.1 – cas bizarre!**
  français → anglais = **64.0**
  français → espagnol = **60.9**
  espagnol → anglais = **60.0**
  anglais → espagnol = **55.2**
  allemand → anglais = **53.6**
  espagnol → français = **51.3**
  anglais → français = **50.1**

- L'anglais se comporte plus comme une langue romane qu'une langue germanique (les scores anglais ↔ {italien,portugais} sont aussi très élevés. Pourquoi?

NRC·CNRC
Institute for Information Technology

Les arbres généalogiques des langues sont simplistes

Les langues se métissent …



NRC·CNRC
Institute for Information Technology

L'histoire triste de l'anglais

- 0-350 AD: les habitants de l'île britannique parlent des langues celtiques, sauf pour une élite qui est bilingue et parle aussi le latin.
- 400-600 AD: les invasions des tribus germanophones qui s'appellent les Angles et les Saxons commencent; ils parlent l'anglo-saxon (dialecte germanique occidental). Les langues celtiques disparaissent de la partie sud-est de l'île, maintenant appellée « England » = « Angeln + Land ».
- 800-1000 AD: les invasions des Vikings, qui parlent le norrois (dialecte germanique nordique). Ils donnent à l'anglo-saxon deux cadeaux empoisonnés – les deux phonèmes « th » (thing, that) – le pronom « they » et 1000 autres mots.
- 1066 AD: avec la conquête de l'Angleterre par Guillaume le Conquérant de Normandie, 350 ans d'oppression linguistique par le français commencent. L'aristocratie anglo-saxonne est extirpée. Au début de cette période, la nouvelle aristocratie ne parle que le français (et parfois le latin). Les paysans opprimés continuent à parler leur langue, qui est stigmatisée et qui ne joue aucun rôle dans le court, ni dans le gouvernement, ni dans l'église (latin). Le « français de Londres » devient un dialecte connu de la langue française.
- 1200-1410 AD: les maîtres francophones prennent des maîtresses anglo-saxonnes … les serviteurs loyaux d'origine anglo-saxonne sont promus … les prêtres utilisent du jargon latin en parlant l'anglo-saxon … une nouvelle langue mixte, l'anglais fait surface.
- 1413 AD: Henry V, d'ascendance française mais ennemi juré de la France, devient roi d'Angleterre. Pour la première fois, l'anglais devient la langue officielle du royaume.

NRC·CNRC Institute for Information Technology

L'anglais: langue germanique ou romane?

français	anglais	allemand
homme	**man**	Mann
main	**hand**	Hand
bras	**arm**	Arm
porc *(l'animal)*	**pig**, **swine**	Schwein
veau *(l'animal)*	**calf**	Kalb
roi	**king**	König
royal	**royal**, **kingly** *(rare)*	königlich
loi	**law**	Gesetz
court *(l'endroit)*	**court**	Hof
porc *(la viande)*	**pork**	Schweinefleisch
veau *(la viande)*	**veal**	Kalbfleisch

En anglais, les mots de base sont germaniques

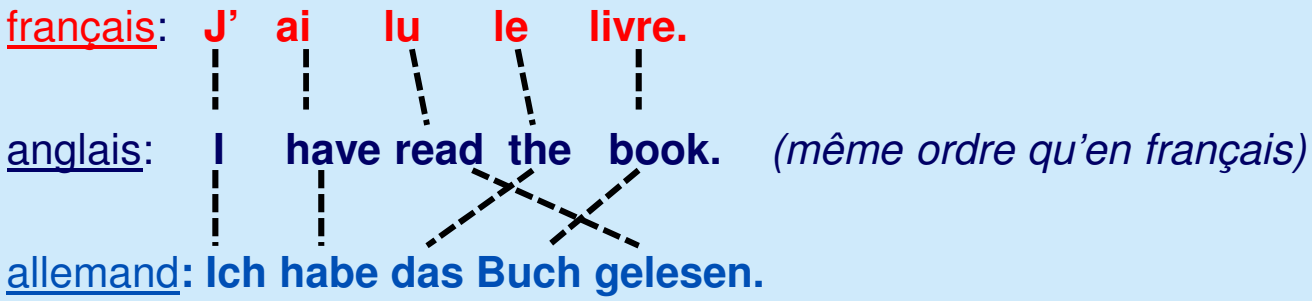
Les mots sophistiqués ou gouvernementaux sont romans

NOTE: 99% des mots dans le « Oxford English Dictionary » **ne sont pas** d'origine anglo-saxonne – mais si on compte chaque occurrence d'un mot dans la langue anglaise parlée, 62% des mots **sont** d'origine anglo-saxonne (« The Power of Babel », J. McWhorter).

NRC·CNRC Institute for Information Technology

L'histoire triste de l'anglais

- Le syntaxe et l'ordre de mots en anglais: fortement influencés par le français.



- **Conclusion**: Ce n'est pas surprenant que l'anglais « se comporte comme une langue romane » - de plusieurs points de vue, l'anglais est une langue romane (bien sûr, une langue romane très atypique!)

NRC·CNRC
Institute for Information Technology

Les facteurs qui prédisent la qualité de la TAS entre deux langues

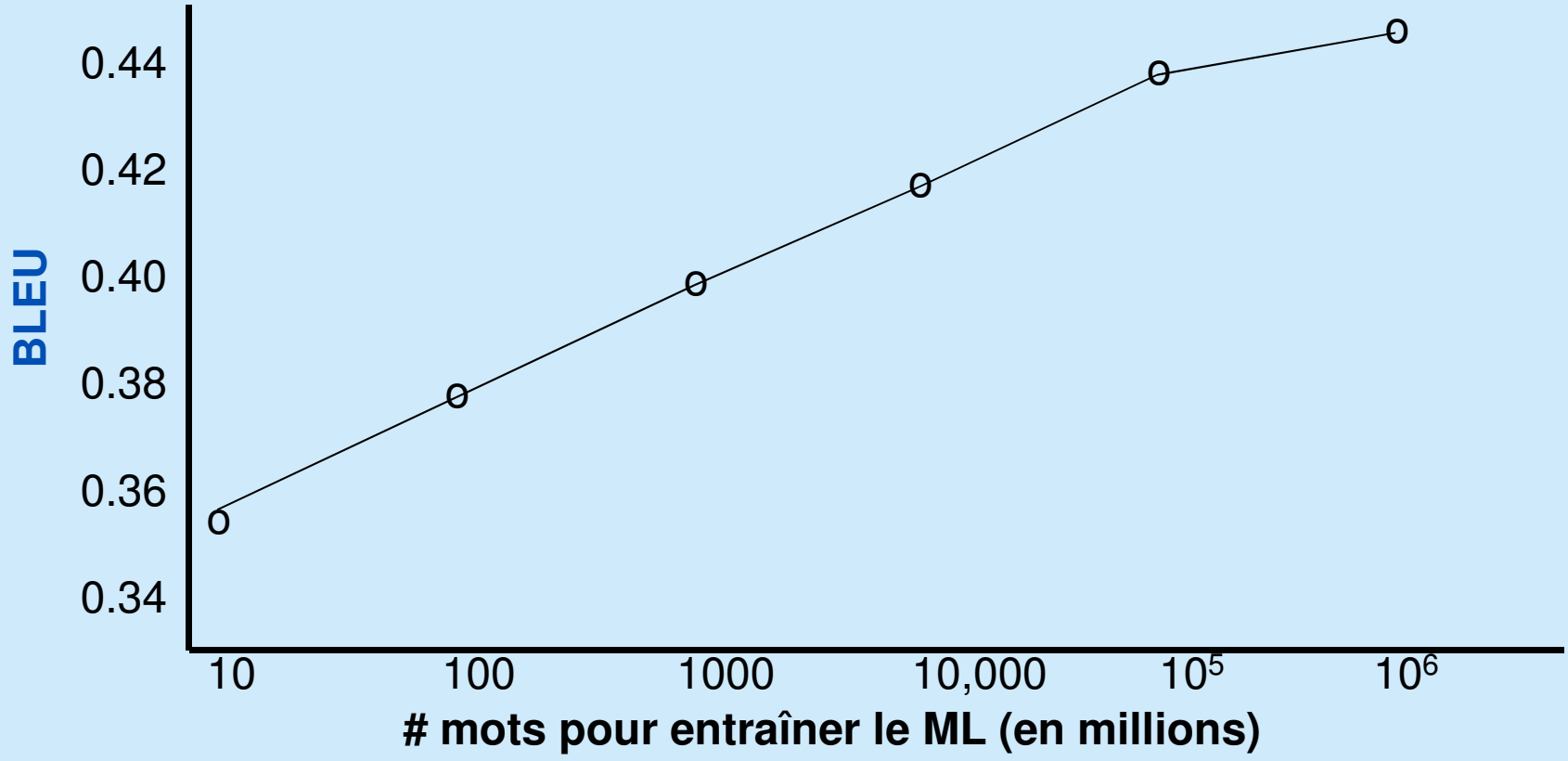
Koehn *et al.* (2009): Basé sur leur analyse de 462 paires de langues européennes, 4 facteurs rendent une paire de langues difficile pour la TAS –

1. La distance généalogique (déjà discutée)
2. La quantité de réordonnancement des mots entre les deux langues
3. Complexité de la morphologie de la langue cible – il est plus difficile de prédire les mots d'une langue qui a une morphologie complexe. Le français a une morphologie plus complexe que celle de l'anglais; dans le tableau de Koehn *et al.*, nous avions des BLEU suivants: français →anglais = **64.0**, mais anglais → français = **50.1**
4. L'entropie – quand il y a beaucoup d'incertitude ou plusieurs choix de traductions pour une grande proportion de mots et de segments de la langue source, l'entropie est élevée. Exemple pour français →anglais: le mot « louche ». « louche » → « sinister »? « shady »? « doubtful »? « untrustworthy »? « sordid »? L'entropie est maximale entre deux langues venant de deux cultures très différentes, où les concepts d'une culture n'ont pas d'équivalent dans l'autre.

NRC·CNRC
Institute for Information Technology

L'avenir de la TAS selon Och

- La disponibilité de beaucoup plus de données réglera la plupart des problèmes! Expériences sur un système arabe → anglais, où toutes les composantes ne changent pas sauf pour le modèle de langue (ML), qui est entraîné sur de plus en plus de données:



NRC·CNRC
Institute for Information Technology

L'avenir de la TAS selon moi

Améliorations de la TAS anglais ↔ français (et certaines autres paires « faciles »):

- Une grande proportion des problèmes d'accord sera réglée bientôt avec l'injection d'un peu de **syntaxe** dans la TAS.
- Une partie des problèmes de co-référence sera réglée bientôt avec une **approche statistique** (par ex., la probabilité du segment « le ministre » *vs.* « la ministre » grandit si le dernier prénom mentionné dans le texte est masculin).
- Une partie des problèmes de registre sera **réglée de la même façon**.
- Oui, **la disponibilité de beaucoup plus de données** aura un impact important.

⇨ **Pas nécessaire de changer l'approche actuelle de façon radicale.**

D'autres paires de langues:
Pour certaines paires de langues telles anglais ↔ chinois, anglais ↔ japonais, l'approche actuelle ne peut pas résoudre les problèmes suivants:

- comment trouver le sujet de la phrase?
- comment (chinois → anglais) décider sur le temps du verbe?
- comment traduire certains concepts qui n'existent même pas dans les langues européennes?
- comment trouver de façon efficace des informations au-delà des frontières de la phrase source actuelles?

⇨ **Pour ces paires, il faut changer l'approche actuelle de façon radicale – il ne suffira pas d'introduire la syntaxe, ou de multiplier les données d'entraînement!**

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Quelles sont les qualités désirables pour les deux corpus d'entraînement = le corpus bilingue pour entraîner le modèle de traduction de segments (MT) et le corpus unilingue en langue cible pour entraîner le modèle de langue (ML)? (3 qualités).
  1. Être dans le même domaine que les documents qu'on veut traduire avec le système
  2. Avoir une taille suffisante
  3. Être de bonne qualité

- Est-ce qu'un système de TAS pourrait bénéficier de l'existence de corpus bilingue (langues source et cible) ou unilingue (langue cible) qui ne sont pas dans le domaine d'intérêt? Si oui, comment?
  Oui. L'algorithme MERT nous permet d'utiliser des modèles MT ou ML entraînés sur des corpus qui ne sont pas dans le domaine d'intérêt, en leur donnant des poids plus faibles que les poids associés aux modèles entraînés sur des corpus qui ne sont pas dans le domaine d'intérêt.

- L'entraînement d'un système TAS a trois phases. Lesquelles?
  1. l'entraînement du modèle de langue (ML) (ou de plusieurs ML)
  2 . l'entraînement du modèle de traduction de segments (MT) (ou de plusieurs MT)
  3. MERT: le calcul du poids sur chaque ML et MT.

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Qu'est qui demande plus de ressources informatiques (puissance de calcul, mémoire, *etc.*) – l'entraînement d'un système TAS ou son utilisation pour faire des traductions?
  L'entraînement d'un système TAS demande par loin plus de ressources que l'utilisation du système.

- Nommez trois utilisateurs de la TAS, et décrire l'usage qu'ils font de la TAS.
  (N'importe quels trois de la liste suivante):
  - Le gouvernement américain: surveillance des pays « menaçants » non-anglophones
  - IBM, Microsoft, et d'autres grandes compagnies américaines: surtout la traduction de leurs manuels de l'anglais vers d'autres langues
  - la Communauté européenne: plus grand consommateur de traduction au monde; traduction de documents officiels entre 23 langues. Une partie de ces traductions est faite par la TAS.
  - Language Weaver: technologie de TAS sur le site du client.
  - PORTAGE: système TAS du CNRC; disponible sous licence aux clients surtout canadiens.
  - Google: service gratuit de TAS disponible à tout le monde via le Web.

- Nommez les sources principales de données utilisées par GoogleTranslate pour entraîner leur systèmes TAS.
  Le « Linguistic Data Consortium » (banque de données académiques, ÉU); les Nations-Unies (documents dans les 6 langues officielles); surtout, des données multilingues sur le Web

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Nommez deux mesures humaines utilisées pour évaluer la qualité des systèmes de TA, avec une brève description de chacune.
  (N'importe quels deux de la liste suivante)
  - Demander aux juges humains de donner un score à la fidélité et un score à la fluidité de chaque sortie d'un système TA (souvent en consultant une traduction de référence)
  - Donner aux juges humains un groupe de traductions venant de plusieurs systèmes TA pour la même phrase source, et demander aux juges de les ordonner en ordre de qualité (souvent en consultant une traduction de référence)
  - Pour mesurer la fidélité seulement, on peut donner aux juges humains des textes traduits par plusieurs systèmes TA et les faire passer un examen sur le contenu de ces textes (approche DLPT*). Si les lecteurs des traductions produites par système X obtiennent en moyen des meilleures notes sur l'examen que les autres juges, c'est une indice que système X est plus fidèle que les autres systèmes.
  - HTER = Human-mediated Translation Edit Rate. On demande aux juges de corriger la sortie du système TAS (souvent en consultant une traduction de référence). Un système qui demande peu de changements est jugé supérieur à un système qui en demande plusieurs.

- Pourquoi est-ce qu'on utilise des mesures automatiques de la qualité des systèmes TA?
  Les mesures humaines demandent trop de temps et sont trop coûteuses. Par exemple, l'étape « MERT » de l'entraînement d'un système TAS demande des milliers d'évaluations de petites variantes du même système: impossible de faire tout ça avec des juges humains.

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Quelle est la mesure automatique la plus utilisée? Lesquels sont ses avantages et désavantages?
  La mesure automatique de la qualité de la TA la plus utilisée s'appelle « BLEU ». C'est une mesure de similarité entre la sortie de la TA et une ou plusieurs traductions de référence, basée sur le nombre de *n*-grammes qui apparaissent et dans la sortie de la TA, et dans les références. Avantages: rapide, peu couteuse, bon pour MERT, corrélation avec les jugements humains (au niveau des systèmes). Désavantages: la corrélation avec le jugement humain au niveau des phrases individuelles n'est pas bonne; une bonne traduction qui a peu de mots et de *n*-grammes en commun avec les références sera pénalisée.

- En faisant la post-édition de la sortie d'un système de TAS, quels types d'erreurs est-ce que vous risquez d'être obligé de corriger le plus souvent? (Donnez 3 types)
  (N'importe quels trois de la liste suivante)
  - Des fautes d'accord à longue distance
  - Des erreurs de référence à longue distance (par ex., « elle » pour « il » ou vice versa)
  - Problèmes sémantiques: la phrase produite par un système TAS n'a pas de sens
  - Problèmes de registre ou de style (par ex., la traduction est vulgaire quand elle devrait être formelle ou vice versa)

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Donnez un aspect de l'anglais qui montre que l'anglais a des origines germaniques. Certains autres aspects de la langue (qui sont importants pour les systèmes TAS) font qu'elle ressemble encore plus à une langue romane – lesquels?
  Les mots « de base » – pour « homme », « eau », « lait », « pain », « main », *etc.* – de l'anglais sont presque tous germaniques. Par contre, il y a deux aspects importants de l'anglais qui sont importants pour la TAS et qui font en sorte que cette langue ressemble aux langues romanes: 1. les mots « sophistiqués » sont d'origine romane 2. l'ordre des mots dans une phrase ressemble à l'ordre dans les langues romanes.

- Imaginez que vous avez un groupe énorme de documents qui viennent de l'ONU, et qui ont été traduits en les 6 langues officielles de l'ONU. Vous décidez d'utiliser les parties appropriées de ce corpus pour entraîner 4 systèmes TAS. SVP ordonner les 4 systèmes résultants selon vos prédictions de la qualité de leur traductions (#1 = meilleur système, #4 = pire système). **NOTE**: le russe est une langue indo-européenne avec une morphologie complexe (comme celle du français ou de l'espagnol); la morphologie de l'anglais n'est pas très complexe. Les 4 systèmes:
  russe → anglais; **anglais → russe**; **chinois → français**; **français → espagnol**.
  Ici, on a seulement une langue non-indo-européenne: le chinois. La traduction entre une paire des langues de familles différentes est très difficile ⇨ **chinois → français** est la plus difficile de ces paires. Le français et l'espagnol sont dans la même sous-famille des langues indo-européennes ⇨ **français → espagnol** est la plus facile de ces paires. Selon Koehn *et al.* (2009), plus la morphologie de la langue cible est complexe, plus la traduction est difficile. Donc **russe → anglais** est plus facile que **anglais → russe**.
  Conclusion: #1 (meilleur) **français → espagnol** , #2 **russe → anglais** , #3 **anglais → russe,** #4 (pire) **chinois → français**.

NRC·CNRC
Institute for Information Technology

Questions sur la TAS auxquelles il faut savoir répondre (2ième classe de RK)

- Selon Franz-Josef Och, laquelle est la méthode la plus efficace pour améliorer les systèmes TAS?
  Entraîner ces systèmes sur **beaucoup** plus de données.