Université du Québec à Montréal
Doctorat en Informatique Cognitive

# Traduction et translittération des entités nommées pour une paire de langues peu dotée

**Ngoc Tan Le**

26 Septembre 2019

# Plan de présentation

# 1.1. Contexte et motivation

- Croissance phénoménale des informations et des connaissances multilingues via les documents numérisés et dans le Web
  ⇒ **Besoins d'outils** de traitement automatique des langues naturelles (TALN) et de la traduction automatique (TA)

- Diversité des langues avec le plus grand nombre d'internautes

**Top Ten Languages Used in the Web - April 30, 2019**
( Number of Internet Users by Language )

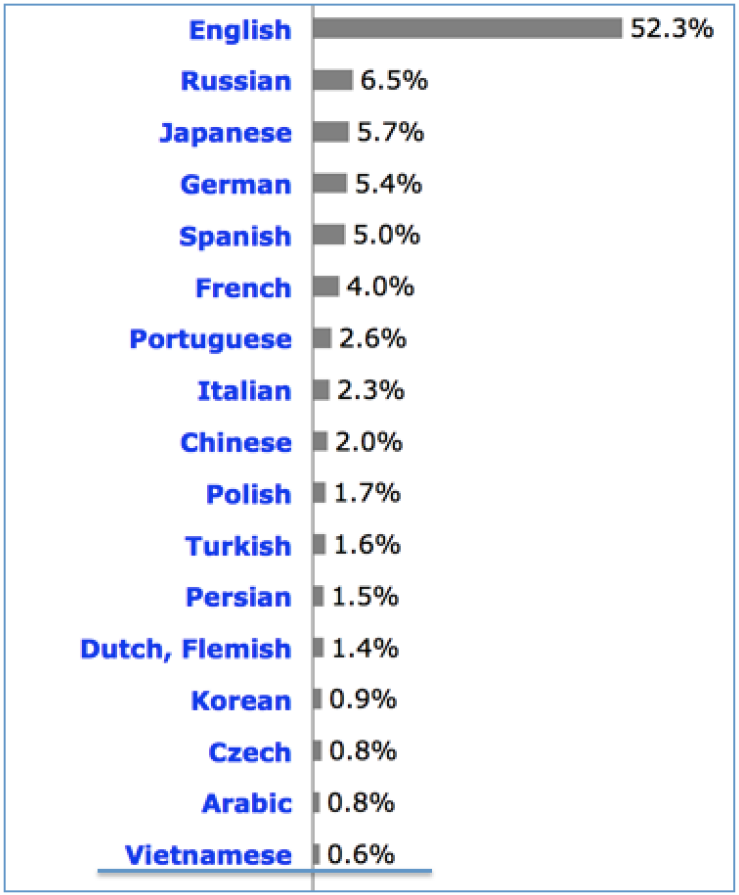
| TOP TEN LANGUAGES IN THE INTERNET | World Population for this Language (2019 Estimate) | Internet Users by Language | Internet Penetration (% Population) | Internet Users Growth (2000 - 2019) | Internet Users % of World (Participation) |
|---|---|---|---|---|---|
| English | 1,485,300,217 | **1,105,919,154** | 74.5 % | 685.7 % | 25.2 % |
| Chinese | **1,457,821,239** | **863,230,794** | 59.2 % | 2,572.3 % | 19.3 % |
| Spanish | 520,777,464 | **344,448,932** | 66.1 % | 1,425.8 % | 7.9 % |
| Arabic | 444,016,517 | **226,595,470** | 51.0 % | 8,917.3 % | 5.2 % |
| Portuguese | 289,923,583 | **171,583,004** | 59.2 % | 2,164.8 % | 3.9 % |
| Indonesian / Malaysian | **302,430,273** | **169,685,798** | 56.1 % | 2,861.4 % | 3.9 % |
| French | 422,308,112 | **144,695,288** | 34.3 % | 1,106.0 % | 3.3 % |
| Japanese | 126,854,745 | **118,626,672** | 93.5 % | 152.0 % | 2.7 % |
| Russian | 143,895,551 | **109,552,842** | 76.1 % | 3,434.0 % | 2.5 % |
| German | 97,025,201 | **92,304,792** | 95.1 % | 235.4 % | 2.1 % |
| TOP 10 LANGUAGES | 5,193,327,701 | **3,346,642,747** | 64.4 % | 1,123.0 % | 76.3 % |
| Rest of the Languages | 2,522,895,508 | **1,039,842,794** | 41.2 % | 1,090.4 % | 23.7 % |
| WORLD TOTAL | 7,716,223,209 | **4,386,485,541** | 56.8 % | 1,115.1 % | 100.0 % |

Source : https ://www.internetworldstats.com/stats7.htm

# 1.1. Contexte et motivation

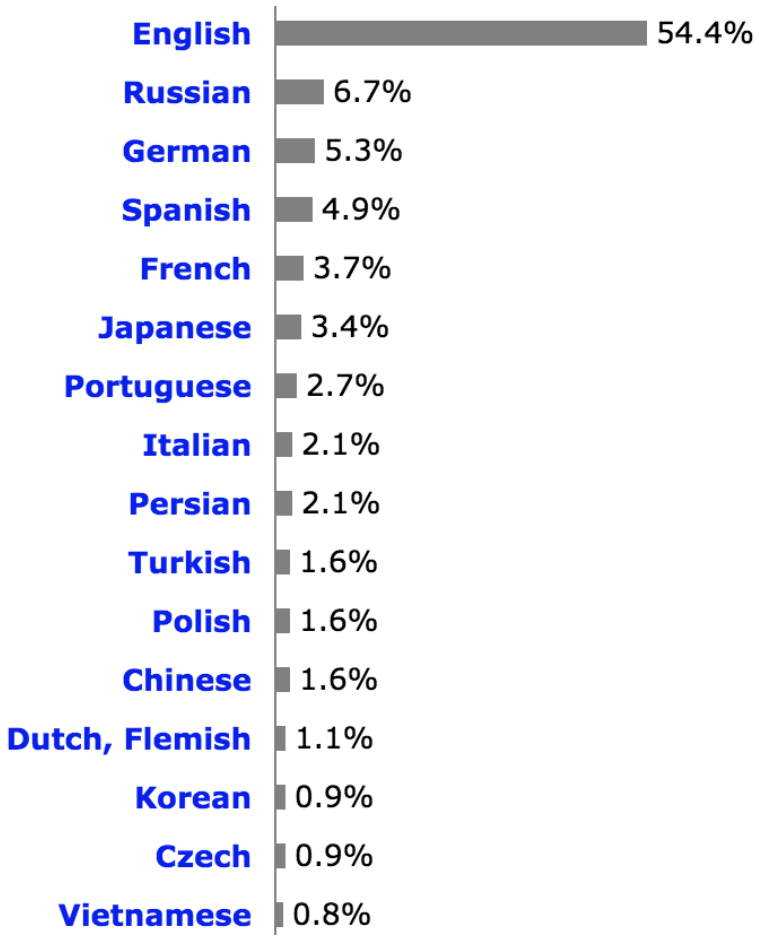
- Diversité des langues les plus utilisées dans les sites Web

Année 2017



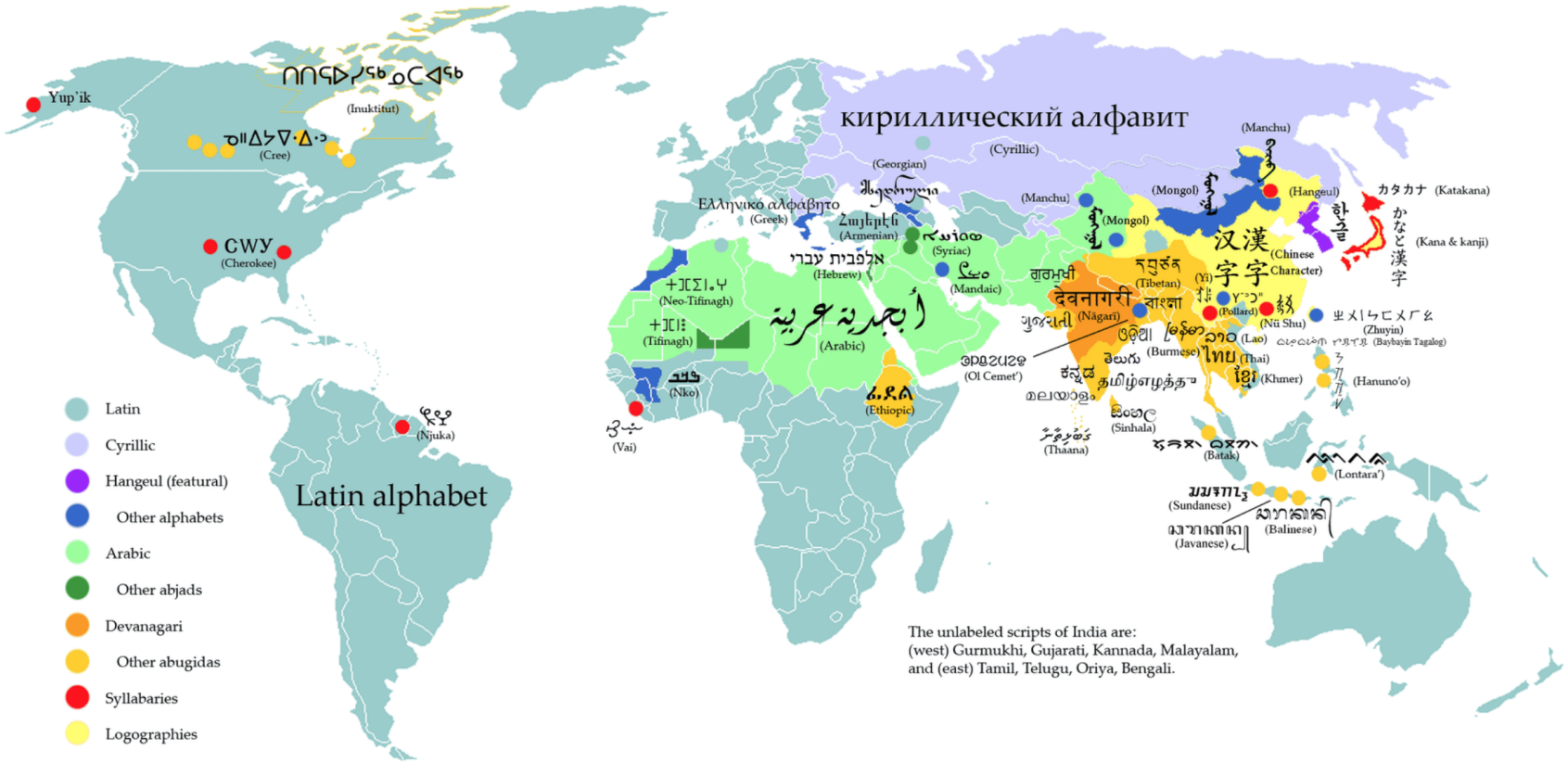

Langues les plus utilisées par les sites d'Internet par Web Technology Surveys 2017
(Source : https://w3techs.com/technologies/overview/content_language/all)

Année 2019



# 1.1. Contexte et motivation

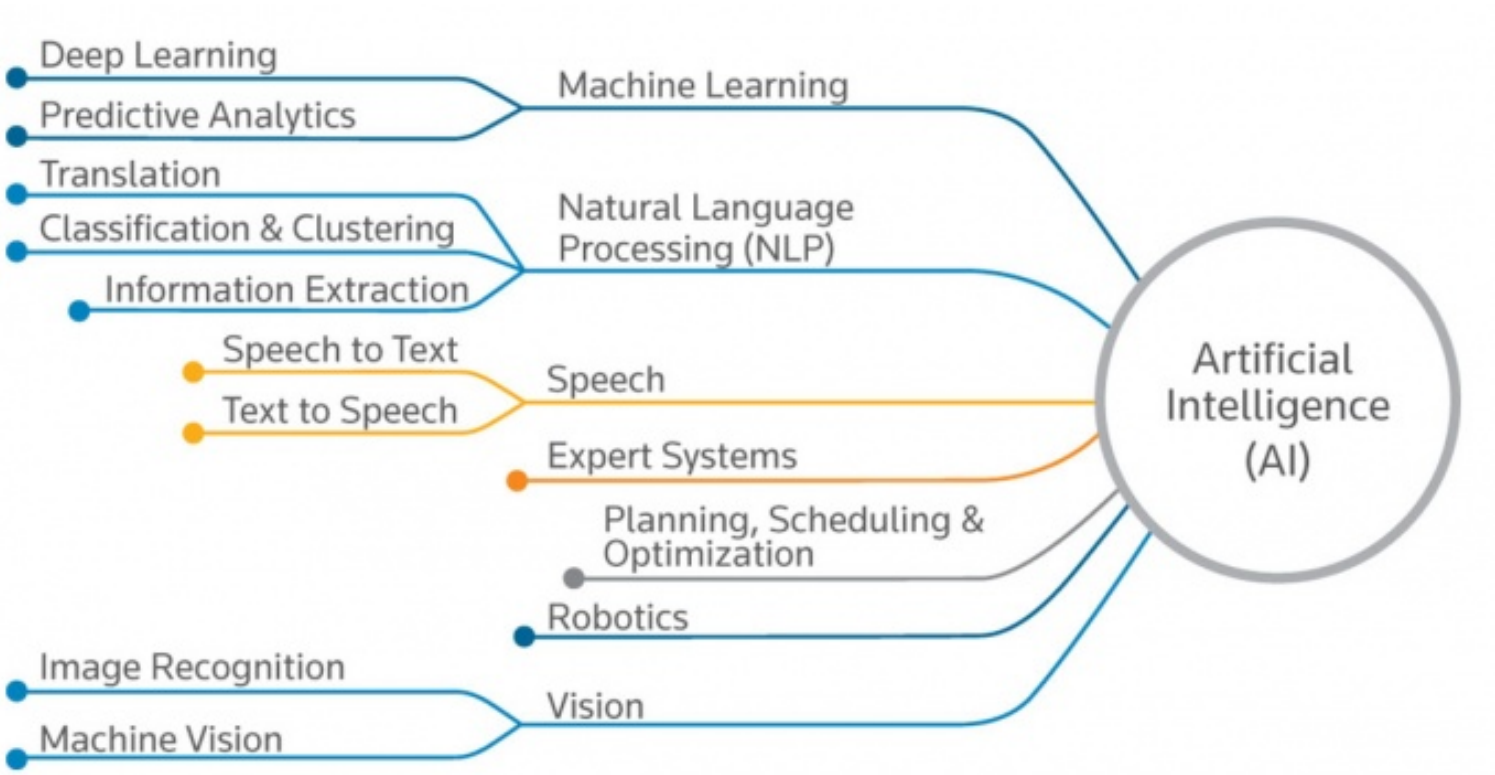
- Diversité des systèmes d'écriture




(Source : Wikipedia, 2019)

# 1.1. Contexte et motivation

- Traduction automatique (TA) : un sous-domaine de la linguistique computationnelle, TALN et Intelligence Artificielle.




Source : https ://blogs.thomsonreuters.com/

# 1.1. Contexte et motivation

- Traduction automatique (TA) : un sous-domaine de la linguistique computationnelle, TALN et Intelligence Artificielle.
  - **Problématiques :**
    - (A) **Les systèmes de TA ne proposent pas toujours de bonnes traductions**
      1. Traduire *un nom commun écrit avec des majuscules comme un nom propre*
      2. Traduire *un nom propre ayant une signification dans un dictionnaire bilingue comme un nom commun*
    - (B) **Mots hors vocabulaires (MHV) ou non traduits** : entités nommées, expressions multi-mots et acronymes.
      etc.

# 1.1. Contexte et motivation

## Définition (Ehrmann, 2008)

*Étant donné un modèle applicatif et un corpus, on appelle « **entité nommée** » toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.*

- Quelques illustrations de la traduction automatique des entités nommées (TAEN)
  (1.1) Hier soir, j'ai mangé avec **Monsieur Michel Poulet**.
  (1.2) Đêm qua tôi đã ăn **thịt gà với Michel**.
  (*traduction erronée, par Google Translate, consultée Lundi 07 Janvier 2018 ; Lundi 06 Mai 2019.*
  *» Littéralement : Hier soir, j'ai mangé* **du poulet avec Michel.**)
  (1.3) Tối qua, tôi đã ăn với **ông Mi-sên Pu-lê**. (*traduction et translittération correcte*)

  (2.1) Ma famille voyage dans **le delta du fleuve Mékong**.
  (2.2) Gia đình tôi đi du lịch ở **đồng bằng sông Mêkông**.
  (*translittération correcte*)
  (2.3) Gia đình tôi đi du lịch ở **đồng bằng sông Cửu Long**.
  (*traduction correcte, par Google Translate, consultée Lundi 07 Janvier 2018 ; Lundi 06 Mai 2019*)

# 1.1. Contexte et motivation

- **Peu de recherches pour une paire de langues peu dotées français-vietnamien**
- **Une langue peu dotée ou langue-$\pi$** : dans le contexte du TALN, peu de ressources linguistiques numériques, d'outils linguistiques (Berment, 2004 ; Besacier et al., 2014)
- **Emprunts des mots** et des **constructions grammaticales** du **chinois**, du **français** et de l'**anglais** au long de l'Histoire du Vietnam

# 1.2. Caractéristiques du vietnamien

- **Alphabet** : *29 lettres*

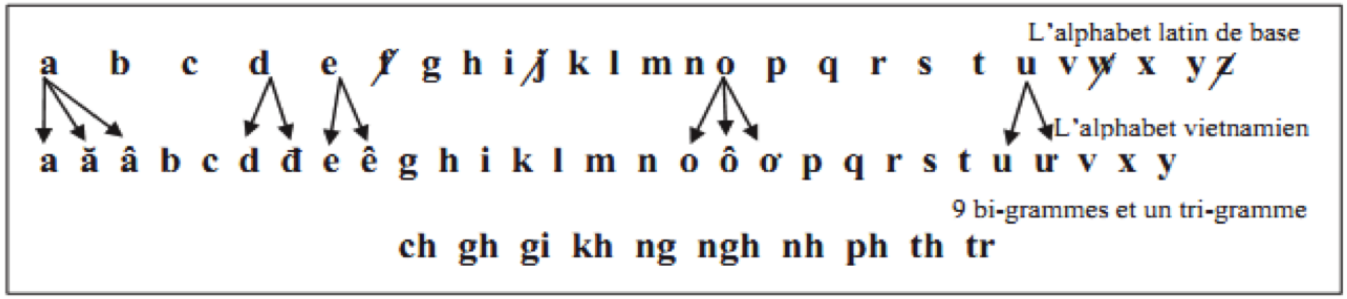



**Figure 3-2 : L'alphabet vietnamien**

(*Source : Do Thi Ngoc Diep (2011). Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée, Thèse 2011, Université de Grenoble*)

- **6 tons** (*ngang, sắc, huyền, hỏi, ngã, nặng*)

**Tableau 3-2 : La signification d'un mot vietnamien dépend de son ton (src : [Tran D.D, 2007])**

| **Mot vietnamien** | *bá* | *bã* | *ba* | *bả* | *bà* | *bạ* |
|---|---|---|---|---|---|---|
| **Signification** | le roi | le marc (de café) | trois | la pâture | grand-mère | n'importe |

# 1.2. Caractéristiques du vietnamien

- Type de langue **isolante** : **non-flexionnelle**, **monosyllabique** (dont *le chinois, le khmer, le thai*)
  - Au niveau morphologique : *Pas de changement*
    *Par exemple :*
    *Tôi nhìn anh ấy = Je regarde lui (littéral) // Je le regarde (correct)*
    *Anh ấy nhìn tôi = Il regarde moi (littéral) // Il me regarde (correct)*
  - Au niveau grammatical : *Ordre des mots*
    *Par exemple :*
    *quần áo // áo quần = le vêtement*
    *Gạo xay = Riz moulu // Xay gạo = Moudre du riz*
    *Nó bảo sao không tới = Il dit pourquoi ne pas venir*
    *Sao không bảo nó tới = Pourquoi ne pas lui dire de venir*
    *Sao bảo nó không tới = Pourquoi dire qu'il ne vient pas*
- **Syllabe** : environ 10 000 syllabes en vietnamien (Hoang Phe, 1997)
  *Par exemple : Tôi / nhìn / anh / ấy ⇒ Il y a 4 syllabes.*
- **Mot** : se compose d'une syllabe ou de plusieurs syllabes
  *Par exemple : Tôi / nhìn / anh_ấy ⇒ Il y a 3 mots.*
- **Problématique** : **Les frontières d'un mot ne peuvent pas être déterminées par un espace**

# 1.3. Traduction/Translittération des ENs en vietnamien

- Noms propres (NPs) en vietnamien : 2 catégories
  (a) NPs d'origine vietnamienne
  *Par exemple : Trần Văn Tạo, Đinh Tiên Hoàng, Lý Bạch, Lỗ Tấn*
  (b) NPs d'origine étrangère
  *Par exemple : François, Ivan, Napoléon, Hugo (en français)*
  **Sans traits d'union** *: Frăngxoa/Phơrăngxoa/Frânxoa/Phơrânxoa, Ivan/Ivân, Napôlêông, Huygô/Uygô*
  **Avec traits d'union** *: Frăng-xoa/Phơ-răng-xoa/Frân-xoa/Phơ-rân-xoa, I-van/I-vân, Na-pô-lê-ông, Huy-gô/Uy-gô*
- Écriture des noms de personnes, de locations, d'organisations :
  - **Influence sino-vietnamienne**
  - **Translittération en sino-vietnamien ou via une langue indo-européenne**

  *Par exemple : Washington, Berlin, Moscou, Tokyo (en français)*
  **Sans traits d'union** *: Oasinhtơn, Béclin/Béclanh, Mátxcơva, Tôkyô*
  **Avec traits d'union** *: Oa-sinh-tơn, Béc-lin/Béc-lanh, Mát-xcơ-va, Tô-ky-ô*
- **Problématiques : Manque de l'uniformité, influences des dialectes**

# 1.4. Objectifs de recherche

- Pourquoi choisir les entités nommées à traiter ?
  - **Emprunts des mots** et des **constructions grammaticales** du **chinois**, du **français** et de l'**anglais**
  - Évolution du Web et mondialisation de l'économie : Une grande diversité des langues de communication ⇒ **Un grand nombre de nouveaux mots**
  - **Mots hors vocabulaire** : entités nommées (**personne, location et organisation**)
- Pourquoi choisir l'apprentissage machine (approche statistique) ?
  - **Avoir la capacité d'apprendre automatiquement les connaissances linguistiques** au niveau lexical et syntaxique
  - Entraîner des systèmes de REN grâce aux **corpus annotés d'apprentissage**
- Pourquoi choisir l'approche à base de réseaux de neurones ?
  - **Émergence** de l'approche à base de réseaux de neurones
  - **Impacts** de l'approche à base de réseaux de neurones dans le traitement des entités nommées

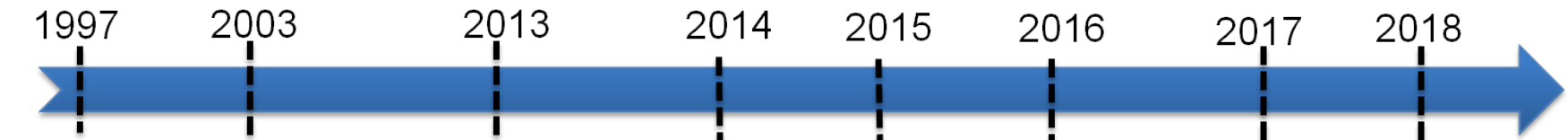
# 2.1. Reconnaissance des entités nommées (REN)

## Définition (Ehrmann, 2008)

*Étant donné un modèle applicatif et un corpus, on appelle « **entité nommée** » toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.*

- La tâche de REN : la première fois dans la 6ème Message Understanding Conference - **MUC** (Sundheim, 1995), **CoNLL**-2003 (Tjong Kim Sang et al., 2003)
- La notion d'EN couvre non seulement les noms propres, mais aussi des entités plus complexes :
  - des noms désignant des **PER**sonnes, **LOC**ations, **ORG**anisations, etc.
  - des notions plus techniques comme les maladies, les gènes (*BioInformatique*)
  - des expressions multi-mots : idiomes, groupes verbaux, groupes nominaux, etc.
- Un grand impact envers le **TALN** :
  - CLIR, IE, QA, **Traduction automatique**
  - Moteurs de recherche, filtrage du contenu Web

# 2.2. Traduction neuronale : Bref historique




Encoder/Decoder (RNN) (Neco et al., 1997)

Neural Language Model (UDeM) (Bengio et al., 2003)

•Encoder-Decoder (CNN+ RNN) (Kalchbrenner and Blunsom., 2013)

• Modèle d'alignement mot à mot dans NMT (Yang et al., 2013)

•Modèle (Seq-to-seq) (Sutskever et al., 2014)

• RNN Encoder–Decoder dans SMT (cho et al., 2014)

•Encodeur/décodeur avec un mécanisme d'attention (Bahdanau et al., 2015)

•Méthode de désambiguisation (OOV) (Jean et al., 2015) UDeM

•Premier traducteur neuronal de Google (Deep RNN) (Wu et al., 2016)

• Dual Learning (Xia et al., 2016)

•Transfer Learning Pour NMT (Zoph et al., 2016)

•Premier TAN (Ar-Ang) (Almahairiet et al., 2016)

•Google Zero-Shot NMT (Universal) (Gu et al., 2017)

• Premier traducteur neuronal de Facebook (CNN) (Wu et al., 2016)

•TAN sans corpus parallèles (Artetxe et al., 2017 et Lample et al., 2017)

•Transfer Learning Pour NMT (Nguyen et Chiang 2017 et Karakanta et al., 2017)

# 2.2. Google Deep NMT

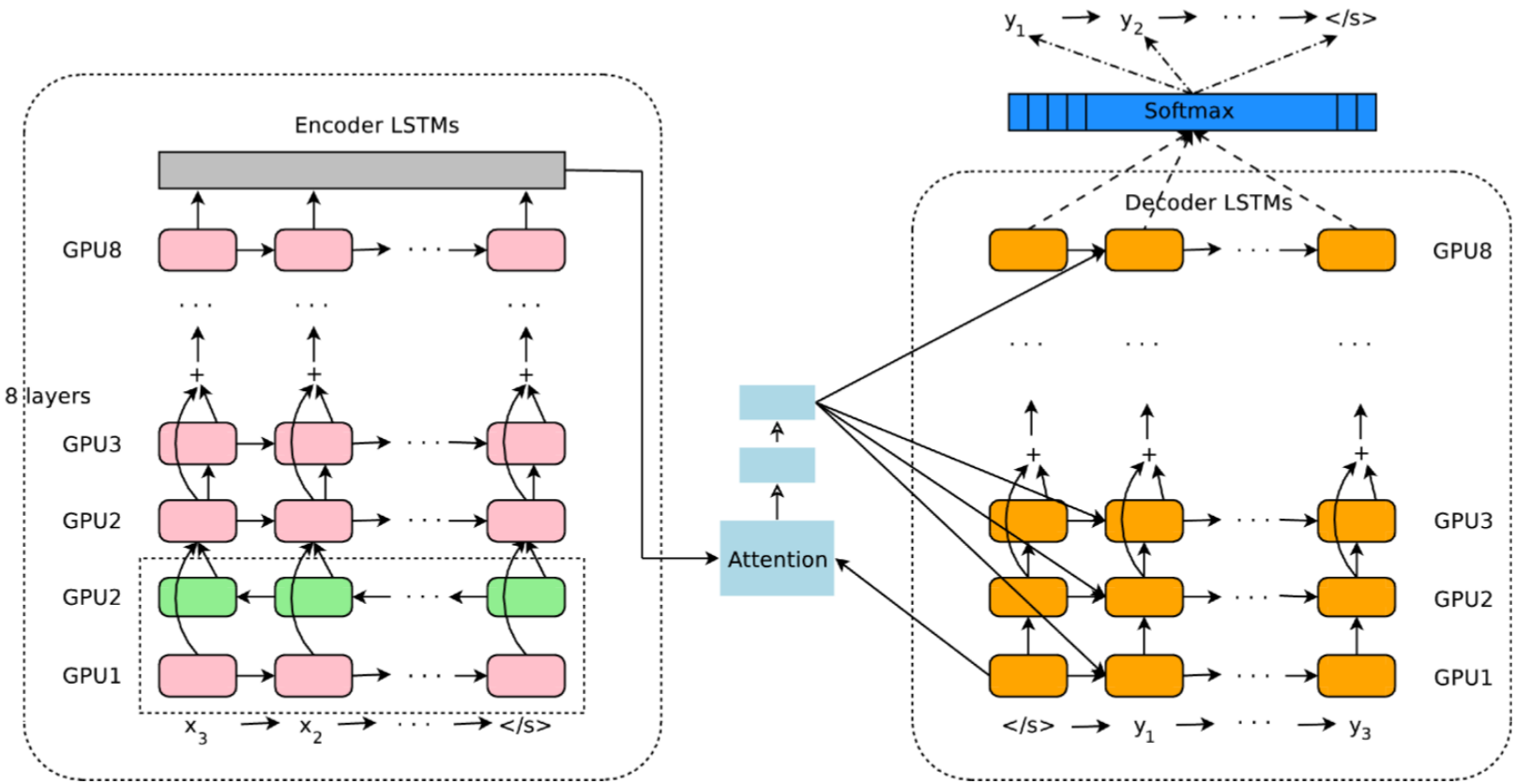



FIGURE – "Google's Multilingual Neural Machine Translation System : Enabling Zero-Shot Translation" (Johnson et al., 2018)

# 3.1. Méthodologie de recherche

- Répartir en tâches suivantes :

1. Collecter des ressources linguistiques à partir des sites Web multilingues
2. Développer un modèle de reconnaissance des entités nommées pour une langue peu dotée de ressources linguistiques
3. Appliquer l'approche à base de réseaux de neurones dans le traitement des entités nommées pour améliorer davantage la qualité et l'efficacité du système
4. Évaluer notre méthodologie

# 3.2. Processus de validation des résultats

- Pour pallier le problème d'indisponibilité de corpus bilingues d'apprentissage
  - Collecter des textes monolingues *à partir des sites Web en vietnamien* (i.e. `http://fr.vietnamplus.vn`, `http://www.vietnamplus.vn`, `http://tamdaoconf.com`)
  - Collecter des *données d'apprentissage* fournies par les campagnes d'évaluation (*i.e. MUC, CoNLL, NEWS*)
- **Métriques d'évaluation**
  - Traitement des entités nommées bilingues : **Précision, Rappel** et **F-score**
  - Traduction automatique
    - Évaluation automatique : score de **BLEU** - *BiLingual Evaluation Understudy* (Papineni et al., 2002)
    - Métrique pour l'évaluation de la traduction avec ordonnancement explicite : **METEOR** - *Metric for Evaluation of Translation with Explicit ORdering* (Banerjee et Lavie, 2005)
    - Taux d'erreur de traduction : **TER** - *Translation Error Rate* (Snover et al., 2009)
    - Taux des mots hors vocabulaires (**MHV**)

# 3.3. Système de traduction automatique statistique des entités nommées (TAEN)

## Le processus global : trois étapes

1. **Prétraitement**
2. **Construction d'un système de TAEN à base de segments** pour une paire de langues peu dotée
3. **Post-traitement**

*(Le et al., ACM TALLIP 2019)*

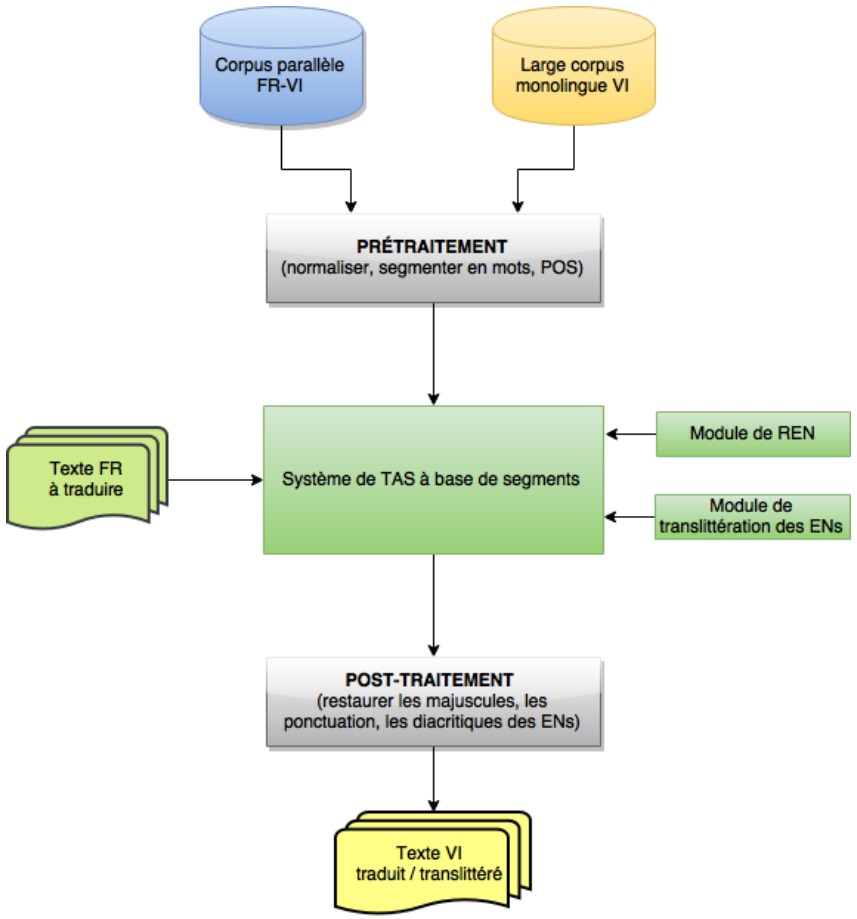



FIGURE – Architecture de notre système de TAEN

# 4.1.1. Module de reconnaissance des entités nommées

**Le processus global : trois étapes**

1. **Pré-traitement** : normaliser le corpus monolingue, segmenter des mots et faire un étiquetage morpho-syntaxique
2. **Extraction des traits linguistiques** : orthographique, information contextuelle locale, etc.
3. **Construction du modèle**

(Le et al., NAACL HLT 2018b)

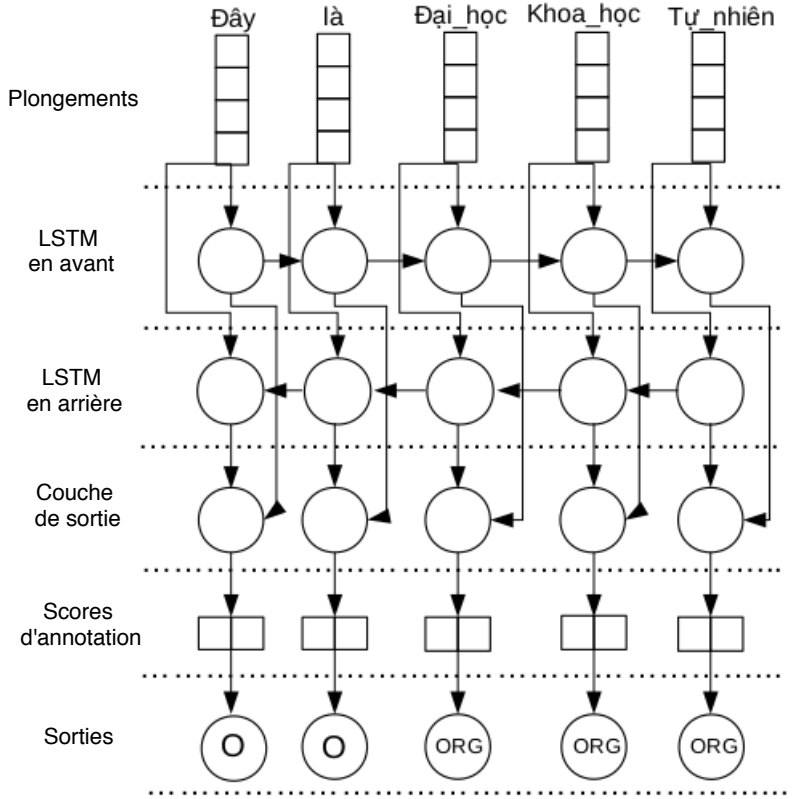



FIGURE – Architecture du modèle de reconnaissance des entités nommées en vietnamien (*Traduction de l'illustration en français : «Voici l'université des sciences naturelles»*)

# 4.1.2. Extraction des traits linguistiques (RNN)

- Illustration de différents traits linguistiques dans une phrase en vietnamien

TABLE – Illustration des traits linguistiques, à savoir surface de mot, lemme, grappe de mot (*word cluster*), étiquette morpho-syntaxique, *chunk* (segment syntaxique) et ENs (entités nommées). Ici, *en vietnamien : « Chỉ_có Pháp và Anh ủng_hộ đề_xuất của Fischler. », en français : « Seules la France et la Grande-Bretagne ont soutenu la proposition de Fischler. »*

| Surface de mot | Lemme | Grappe de mot | Étiquette POS | Chunk | ENs |
|---|---|---|---|---|---|
| Chỉ_có | chỉ_có | 011000111 | RB | I-NP | O |
| Pháp | pháp | 111001 | NNP | I-NP | I-LOC |
| và | và | 0011110 | CC | I-NP | O |
| Anh | anh | 111001 | NNP | I-NP | I-LOC |
| ủng_hộ | ủng_hộ | 0110011011 | VBD | I-VP | O |
| đề_xuất | đề_xuất | 01010111 | NN | B-NP | O |
| của | của | 001110 | POS | I-NP | O |
| Fischler | fischler | 0110100 | NNP | I-NP | I-PER |
| . | . | 0010 | . | O | O |

# 4.1.3. Expérimentations

- Préparation des données :

| | Langue | Catégorie | #App. (tokens) | #Dév. | #Test |
|---|---|---|---|---|---|
| **WNUT'2016** | anglais | tweets | 2 394 (37 619) | 420 | 3 856 |
| **CAP'2017** | français | tweets | 3 000 (62 261) | null | 3 000 |
| **CLC'2018** | vietnamien | actualité | 9 600 (248k) | 1 200 | 1 200 |

TABLE – Statistiques des données d'apprentissage

  - Un large corpus monolingue non étiqueté, contenant 17 095 994 phrases en vietnamien (équivalent à 372 720 876 *tokens*)
- Outils utilisés :
  - Segmenteur au niveau des mots CLC_VN_WS [1], *word2vec* [2], *TreeTagger* [3], *conlleval* [4] script : évaluer les modèles de REN
  - Implémentations de Wapiti-CRF [5] et Bi-LSTM-CRF NER [6] (Lample et al., 2016)

1. http ://www.clc.hcmus.edu.vn/
2. https://code.google.com/p/word2vec/
3. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
4. http ://www.cnts.ua.ac.be/conll2003/ner/
5. https://wapiti.limsi.fr/
6. https://github.com/glample/tagger/

# 4.1.4. Expérimentations

- Les catégories des entités nommées à traiter

| Catégorie | App. | Dév |
|---|---|---|
| person | 266 | 664 |
| geo-location | 158 | 325 |
| company | 49 | 207 |
| facility | 77 | 209 |
| product | 158 | 177 |
| music artist | 76 | 116 |
| movie | 30 | 80 |
| sports team | 83 | 74 |
| tv show | 2 | 65 |
| other | 229 | 545 |
| **Total** | **1 128** | **2 462** |

TABLE – Statistiques des types d'entités nommées pour WNUT'2016, *tweets* en anglais.

| Catégorie | #EN | Catégorie | #EN |
|---|---|---|---|
| musicartist | 80 | transportLine | 548 |
| product | 340 | event | 161 |
| person | 966 | facility | 329 |
| geoloc | 700 | sportsteam | 87 |
| movie | 40 | tvshow | 47 |
| other | 279 | org | 358 |
| media | 191 | **Total** | **4 126** |

TABLE – Statistiques des types d'entités nommées pour CAp'2016, *tweets* en français.

# 4.1.4. Expérimentations

- Les catégories des entités nommées à traiter

| Catégorie | EN |
|---|---|
| ABB | 1 137 |
| ABB_DES | 84 |
| ABB_LOC | 372 |
| ABB_ORG | 686 |
| ABB_TRM | 69 |
| ABB_TTL | 8 |
| BRN | 220 |
| DES | 1 029 |

| Catégorie | EN |
|---|---|
| DTM | 2 684 |
| LOC | 4 299 |
| MEA | 2 618 |
| NUM | 4 754 |
| ORG | 3 279 |
| PER | 4 161 |
| TRM | 190 |
| TTL | 988 |

TABLE – Statistiques du corpus annoté en vietnamien.

Ce corpus contient **26 578** d'entités nommées : Person (PER), Organization (ORG), Location (LOC), Abbreviation (ABB), Brand (BRN), Designation (DES), Date time (DTM), Measurement (MEA), Terminology (TRM), Title (TTL).

# 4.1.4. Expérimentations - Configuration

- Approche à base de réseaux de neurones :

**Algorithme** : Recurrent neural network (RNN)
**Outil** : bi-LSTM-CRF (Lample et al., 2016)
**Optimisation** : sgd (stochastic gradient descent) (Rumellhart et al., 1986)
**Taux d'apprentissage (learning rate)** : 0,01
**Taille des échantillons (batch-size)** : 32
**Nombre d'itération (#epoch)** : 100
**Taux d'abandon (drop-out rate)** : 0,5
**Taille du plongement des mots (embedding) :** 100

# 4.1.5. Résultats obtenus - Tweets

- Twitter en anglais

| | P | R | F1 |
|---|---|---|---|
| **Baseline** | 26,80 | 15,73 | 19,82 |
| **+ Words clusters** | 34,73 | 26,36 | 29,97 |
| **+ Gazetteer** | 41,90 | 27,57 | 33,25 |
| **Notre approche** | 52,66 | 29,63 | **37,92** |

TABLE – Résultats des systèmes de REN avec des *tweets* en anglais. Baseline : avec des traits linguistiques.

- Twitter en français

| | P | R | F1 |
|---|---|---|---|
| **Baseline** | 50,08 | 31,19 | 38,44 |
| **+ Words clusters** | 69,42 | 41,49 | 51,93 |
| **+ Gazetteer** | 69,15 | 41,86 | 52,15 |
| **Notre approche** | 69,46 | 41,94 | **52,30** |

TABLE – Résultats des systèmes de REN avec des *tweets* en français. Baseline : avec des traits linguistiques.

# 4.1.5. Résultats obtenus - Tweets

| Rang | Système | P | R | F1 |
|---|---|---|---|---|
| 1 | CambridgeLTL | 60,77 | 46,07 | **52,41** |
| 2 | Talos | 58,51 | 38,12 | **46,16** |
| 3 | akora | 51,70 | 39,48 | **44,77** |
| 9 | UQAM-NTL | 40,73 | 23,52 | **29,82** |
| | Notre approche | 52,66 | 29,63 | **37,92** |

TABLE – Comparaison de notre modèle avec les trois premiers systèmes participant dans WNUT'2016.

| Rang | Système | P | R | F1 |
|---|---|---|---|---|
| 1 | Synapse Dev | 73,65 | 49,06 | **58,89** |
| 2 | High Inst. of Techno | 58,95 | 46,83 | **52,19** |
| 3 | TanDam | 60,67 | 45,48 | **51,99** |
| 2* | Notre approche | 69,46 | 41,94 | **52,30** |

TABLE – Comparaison de notre modèle avec les trois premiers systèmes participant dans CAp'2017.

# 4.1.5. Résultats obtenus - Corpus en vietnamien

- Expérimentations : 4 systèmes

(1) Système 1 : système de base (*Support Vector Machine*)

(2) Système 2 : système de base (*Conditional Random Fields*)

(3) Système 3 : système de base, dit Bi-LSTM sans les traits linguistiques

(4) Système 4 : notre approche avec les traits linguistiques

TABLE – Comparaison de performances entre différents systèmes de reconnaissance des entités nommées

| Expérimentations | Précision (%) | Rappel (%) | F1 (%) |
|---|---|---|---|
| Système 1 (SVM) | 85,23 | 78,02 | 81,46 |
| Système 2 (CRF) | **86,70** | 79,54 | 82,97 |
| Système 3 (Bi-LSTM, sans traits) | 81,08 | 83,50 | 82,27 |
| Système 4 (notre approche, avec traits) | 84,53 | **87,93** | **86,20** |

# 4.1.5. Évaluations - Prédictions

- Pour le français :
  **Yannick Jadot** en campagne à **Toulouse** pour la primaire présidentielle https ://t.co/Z8t4wzELEQ
  **Yannick_B-person Jadot_I-person** en_O campagne_O à_O **Toulouse_B-geoloc** pour_O la_O primaire_O présidentielle_O https ://t.co/Z8t4wzELEQ_O

- Pour l'anglais :
  '**Breaking Dawn**' Returns to **Vancouver** on January 11th http ://bit.ly/dbDMs8
  '_O **Breaking_B-movie Dawn_I-movie** '_O Returns_O to_O **Vancouver_B-geo-loc** on_O January_O 11th_O http ://bit.ly/dbDMs8_O

- Pour le vietnamienne :
  **Bộ GD-ĐT** cho_biết ngày thi thứ hai của kì thi tốt_nghiệp **THPT năm 2009** có thêm **842** thí_sinh bỏ thi . (*Traduction : "Le ministère de l'Éducation et de la Formation a donné à la deuxième journée d'examen de l'examen de fin d'études secondaires en 2009 plus de 842 candidats."*)
  **Bộ_B-ABB_ORG GD-ĐT_I-ABB_ORG** cho_biết_O ngày_O thi_O thứ__O hai__O của_O kì_O thi_O tốt_nghiệp_O **THPT_B-ABB năm_B-DTM 2009_I-DTM** có_O thêm_O **842_B-NUM** thí_sinh_O bỏ_O thi_O ._O

# 4.1.5. Analyse des erreurs

1. Erreurs de prédiction de nouvelles entités nommées
2. Erreurs de prédiction d'annoter une classe au lieu d'une autre classe

- Pour les messages de Twitter
  - Certaines catégories d'entités nommées : *event*, *media* et *other*.

| Texte | Référence | Prédiction |
|---|---|---|
| Tchat des Jeudis de la Ligne | event | other |
| #TableRondeRATP | event | O |
| #BFM | media | O |

# 4.1.5. Analyse des erreurs

- Pour le vietnamien

TABLE – Illustration des erreurs de prédiction de *nouvelles entités nommées*.

| | | |
|---|---|---|
| Theo Vv O O | | |
| thống_kê Vv O O | 613 X B-NUM B-NUM | |
| của Nn O O | em Nn O O | tại Cm O O |
| riêng Aa O O | dưới Nn O O | **khoa Nn B-ORG O** |
| **Bệnh_viện Nn B-ORG B-ORG** | 18 X B-MEA B-MEA | **Kế_hoạch Nn I-ORG O** |
| **phụ_sản Nn I-ORG I-ORG** | tuổi Nn I-MEA I-MEA | **gia_đình Nn I-ORG O** |
| **Từ_Dũ X I-ORG I-ORG** | đến Vd O O | của Nn O O |
| , PU O O | xin Vv O O | bệnh_viện Nn O O |
| năm Nt B-DTM B-DTM | bỏ Vv O O | . PU O O |
| 2007 X I-DTM I-DTM | thai Nn O O | |
| có Ve O O | | |

*Traduction :*
*Bệnh_viện phụ_sản Từ_Dũ = Hôpital de maternité de Từ_Dũ*
*khoa Kế_hoạch gia_đình = Département de planification familiale*

# 4.2.1. Translittération automatique des entités nommées

- Approche proposée : à base de réseaux neuronaux récurrents
  (1) Pré-traitement
  (2) Modification des séquences d'entrée basée sur la représentation d'alignement
  (3) Création d'un système de translittération à base de réseaux de neurones récurrents

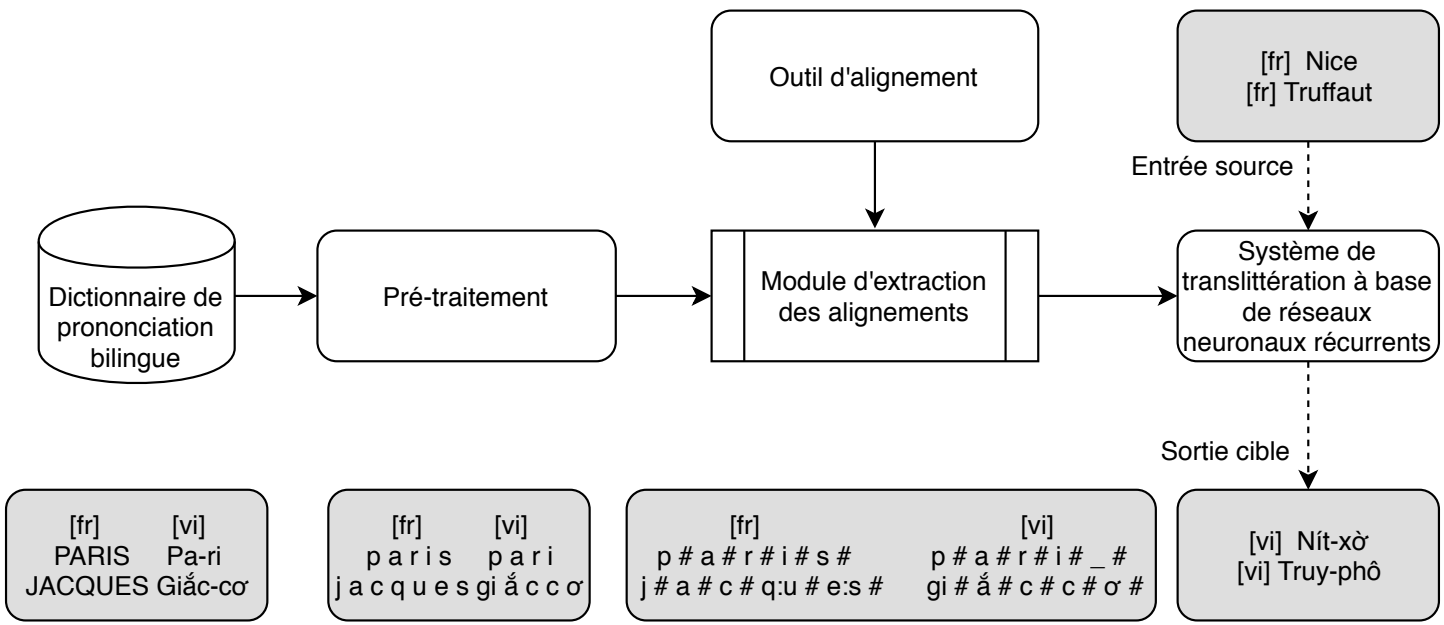



FIGURE – Architecture du module de translittération des entités nommées

(Le et al., ACM TALLIP 2019)

# 4.2.2. Expérimentations - Configuration

- Préparation des données :
  - 4 259 paires d'entités nommées bilingues français-vietnamien (depuis la campagne NEWS [7] - ACL 2018)
- Outils utilisés :
  - *m-2-m aligner* [8] (Jiampojamarn et al., 2007)
  - *word2vec* [9]
  - *nmt-keras* [10] (Peris, 2017) pour entraîner un modèle neuronal de translittération
  - *Sequitur-G2P* [11] pour entraîner un modèle de translittération
  - Moses (Koehn, 2009) [12] pour entraîner un modèle de traduction automatique statistique

7. http://workshop.colips.org/news2018/dataset.html
8. https://github.com/letter-to-phoneme/m2m-aligner/
9. https://code.google.com/archive/p/word2vec/
10. https://github.com/lvapeab/nmt-keras/
11. https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html
12. http://www.statmt.org/moses_steps.html

# 4.2.3. Évaluations

TABLE – Évaluation des expérimentations avec les métriques d'évaluation telles que *BLEU (BiLingual Evaluation Understudy)* (Papineni et al., 2002), *taux d'erreurs de traduction (TER - Translation Error Rate)* (Snover et al., 2009) et *taux d'erreurs de phonèmes (PER - Phoneme Error Rate)*.

| **Expérimentations** | **BLEU ↑** | **TER ↓** | **PER ↓** |
|---|---|---|---|
| Système de base A (pbSMT) | 61,30 | 24,08 | 44,20 |
| Système de base B (modèle de séquences multi-jointes) | 65,70 | 20,50 | 38,00 |
| Système 1 (encodeur-décodeur + attention) | 92,38 | 9,69 | 18,28 |
| Système 2 (encodeur-décodeur + attention + représentation d'alignement) | 94,41 | 4,70 | 8,87 |
| Système 3 (encodeur-décodeur + attention + représentation d'alignement + plongements de sources et de cibles pré-entraînés) | 95,96 | 3,28 | 6,19 |

# 4.2.3. Évaluations

- Prédiction par nos modèles

TABLE – Exemples de résultats de prédiction de translittération par deux systèmes d'évaluation, hypothèses et références pour six noms propres tels que *PARIS, NICE, JACQUES, VINCENTE, YUKON et ZIMBABWE*.

| Évaluation | Hypothèses | Références |
|---|---|---|
| **Système de base A (pbSMT)** | p a r í t<br>n í t<br>gi ắ c<br>v a nh s â n t ơ<br>iu k ô ng<br>gi anh b a b u ê | p a r i<br>n í t x ờ<br>gi ắ c c ơ<br>v a nh x â n t ơ<br>d uy c ô ng<br>gi i m b a b u ê |
| **Notre approche (Système 3)** | p a r i<br>n í t x ờ<br>gi ắ c c ơ<br>v a nh x â n t ơ<br>d uy c ô ng<br>gi i m b a b ơ | p a r i<br>n í t x ờ<br>gi ắ c c ơ<br>v a nh x â n t ơ<br>d uy c ô ng<br>gi i m b a b u ê |

# 4.3.1. Système de traduction et de translittération automatique des entités nommées (TAEN)

- Préparation des données :

TABLE – Statistiques des données d'apprentissage

| **Données** | **Description** | **#paires de phrases** | **#tokens fr - vi** |
|---|---|---|---|
| Apprentissage | conférence Tam Dao 2009 | 5k | **105 009 - 143 255** |
| | conversations, livres bilingues, Web | 15k | **234 422 - 287 178** |
| | corpus utilisé dans la thèse de Diep (2010) | 50k | **1 783 349 - 2 337 997** |

- Outils utilisés :
  - Prétraitement : en vietnamien avec CLC_VN_WS (Vu, 2011)
  - *SRI Language Modeling (SRILM)* (Stolcke et al., 2002),
  - *Moses* (Koehn, 2009) pour entraîner un modèle de traduction automatique statistique
  - *nmt-keras* (Peris, 2017) pour entraîner un modèle de traduction automatique neuronale

# 4.3.2. Système de traduction et de translittération automatique des entités nommées (TAEN)

- Expérimentations :
  1. Expérimentation 1 (Système 1 - TAS) : #app. = 50k, #test = 359 entités extraites de 1060 paires de phrases bilingues
  2. Expérimentation 2 (Système 2 - TAS) : similaire à Exp1, avec modèle de langue entraîné sur un large corpus ayant 17 095 994 phrases en vietnamien.
  3. Expérimentation 3 (Système 3 - TAS) : similaire à Exp2, #app. = 70k
  4. Expérimentation 4 (Système 4 - TAN) : similaire à Exp3

TABLE – Évaluation des performances de différents systèmes de traduction automatique des entités nommées pour la paire de langues peu dotée en français-vietnamien, en appliquant nos deux méthodes.

| Expérimentation | BLEU | METEOR | TER | Taux de MHV |
|---|---|---|---|---|
| Système 1 - TAS (*base*) | 31,40 | 40,50 | 67,60 | 49,80 |
| Système 2 - TAS | 40,00 | 63,40 | 49,70 | 38,60 |
| Système 3 - TAS | **51,80** | **68,40** | **36,20** | **27,50** |
| Système 4 - TAN | 12,04 | 28,30 | 91,94 | 69,72 |

# 4.3.3. Système de traduction et de translittération automatique des entités nommées (TAEN)

- Prédiction par nos modèles

TABLE – Quelques illustrations de prédiction de la traduction automatique des entités nommées produites par les quatre systèmes d'expérimentation, avec un ensemble d'entités nommées en français {*Indonésie*, *Cambodge*, *Asie du Sud-Est*, *Mékong*, *khmer*}.

| **Référence** | **Système 1** | **Système 2** | **Système 3** | **Système 4** |
|---|---|---|---|---|
| inđônêxia | inđônêxia | inđônêxia | inđônêxia | in-đô-nê-xia ta lực |
| campuchia | cam-pu-chia | campuchia | campuchia | tìm thấy miến |
| đông nam á | nam á | đông nam á | đông nam á | nam á thế á thế |
| mêkông | sông | sông mê công | mê công | sông cửu long |
| khmer | dân | khơme | khơme | liệu họ . |

# 4.3.4. Système de traduction et de translittération automatique des entités nommées (TAEN)

- Aller plus loin : Expérimenter notre meilleure approche (Système 3 - TAS)

TABLE – Évaluation de la performance de notre approche pour la tâche de traduction automatique de 100 noms propres de la Bible bilingue française-vietnamienne, ainsi que 100 termes techniques de chimie en général.

| Expérimentation | BLEU | METEOR | TER | Taux de MHV |
|---|---|---|---|---|
| Bible | **42,10 ( +8.90)** | **34,30 ( -3,30)** | **38,80 ( -15,40)** | **58,30 ( -20,30)** |
| Termes techniques | 33,20 | 37,60 | 54,20 | 78,60 |

# 4.3.5. Système de traduction et de translittération automatique des entités nommées (TAEN)

- Prédiction par le système 3 - TAS

- Prédiction de traduction automatique des noms propres, extraits de la bible française-vietnamienne

| Source FR | Référence VI | Prédiction VI |
|---|---|---|
| israel | ysơraên | isờraen |
| ruben | rubên | cuben |
| simeon | simêôn | simêôn |
| levi | lêvi | lêvi |
| juda | giuđa | đuđa |
| issacar | ysaca | ixờtaca |
| zabulon | sabulôn | giabulông |
| benjamin | bêngiamin | bendamin |
| jacob | giacốp | dacốp |
| joseph | giôsep | puydốtxơ |
| ramses | ramse | ramxơ |
| pharaon | pharaôn | pharaông |

- Prédiction de traduction automatique des termes techniques de chimie

| Source FR | Référence VI | Prédiction VI |
|---|---|---|
| adrenaline | ađrênalin | ađờrênalinnơ |
| aleurone | alơron | alêuronnơ |
| amygdale | amiđan | amycđanlơ |
| aminoacide | aminôaxít | aminôaxiđơ |
| amine | amin | aminnơ |
| amibe | amíp | amibê |
| ammoniac | amôniắc | ammôniác |
| acetylene | axêtylen | asờtilânơ |
| blouson | bờludông | bờlôusông |
| hemoglobine | hêmôglôbin | hêmôgờlôbinnơ |
| paludisme | baluyđít | baluđisờme |
| vaseline | vadơlin | vagienlinnơ |
| virus | virút | virútxơ |
| vitamine | vitamin | vitaminnơ |

# Conclusion

- Relever les défis de recherches sur le TALN, spécifiquement **en regard du vietnamien**, *créer les outils linguistiques*

- **Reconnaissance des ENs** et **traduction des ENs** pour le français-vietnamien : **un défi majeur de la TA**
  - *Une grande partie de mots non ou mal traduits dans les textes bilingues en français-vietnamien*

  $\Rightarrow$ **Contribution sur la translittération des ENs bilingues** : *réduire des MHV, non ou mal traduits ainsi qu'améliorer la qualité du système de la TA*

- Notre méthodologie de TAEN : **être adaptée à n'importe quelle paire de langues, à condition de posséder un petit corpus bilingue parallèle**

# Projets concernant le traitement des ENs

- **Projets effectués**
  - (Le et Sadat, ACL - NEWS 2018, Melbourne, Australie) : Translittération des entités nommées en langues asiatiques
  - (Le et Sadat, NAACL - WiNLP 2018, Louisiane, États-Unis) : Reconnaissance des entités nommées en vietnamien
  - (Le et al., CAP compétition 2017, Grenoble, France) : Reconnaissance des entités nommées dans les tweets en français
  - (Le et al., COLING - WNUT compétition, 2016, Osaka, Japon) : Reconnaissance des entités nommées dans les tweets en anglais

- **Projets en cours**
  - Thales Group Québec :
    - Supervision de Fatiha Sadat
    - Approches : bi-LSTM-CRF (Lample et al., NAACL 2016), bi-LSTM-CNN-CRF (Ma et Hovy, ACL 2017)
    - Catégories à traiter : *ORG, USER, DEVICE, APPLICATION, CONNECTION, GAME, SOFTWARE*

# Publications

2019 (a) Low Resource Machine Transliteration Using Recurrent Neural Network. *In Proceedings of ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), ISSN 2375-4699, Inf. Process. 18(2) : 13 :1-13 :14 (2019).*

(b) Augmenting Named Entity Recognition with Commonsense Knowledge. *In Proceedings of ACL 2019, Widening Natural Language Processing workshop, Florence, Italy, 28 July 2019.*

(c) Transfer Learning to Enhance Korean Word Segmentation on Social Media Texts. *In proceedings of the 2019 16th International Conference of the Pacific Association for Computational Linguistics (PACLING), October 11-13, 2019, Hanoi, Vietnam (accepted/accepté).*

# Publications

2018 (a) Cross-linguistic Projection for French-Vietnamese Named Entity Recognition. *Language Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics, Springer LNAI (Lecture Notes in Artificial Intelligence) Series, Lecture Notes in Computer Science, DOI : 10.1007/978-3-319-93782-3_29. In book : Human Language Technology. Challenges for Computer Science and Linguistics, 407-419, June 2018.*

(b) Neural network-based model for Named Entity Recognition in low-resource settings. *In Proceedings of NAACL HLT 2018, Widening Natural Language Processing workshop, New Orleans, USA, 1-6 June 2018.*

2017 (a) Reconnaissance des entités nommées dans les messages Twitter en français. *In Proceedings of 19ème Conférence sur l'apprentissage automatique, Grenoble, France, 28-30 Juin 2017.*

(b) Information Extraction in Real-World Business Documents. *CICLing : International Conference on Computational Linguistics and Intelligent Text Processing, April 17-23, 2017 Budapest, Hungary.*

# Publications

2016 (a) UQAM-NTL : Named entity recognition in Twitter messages. *COLING 2016, Osaka, Japan, 11-16 December 2016.*

(b) Construction de ressources linguistiques pour un corpus annoté bilingue français-vietnamien. *84ème Congrès de l'ACFAS, Université du Québec à Montréal, Québec, Canada, 9-13 Mai 2016.*

2015 (a) Error Analysis of Named Entity Translation output for Poor-Resourced Bilingual Vietnamese-French Pair. *7th Language Technology Conference, Poznan, Poland, 304-308, 27-29 November 2015.*

(b) Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection. *22ème TALN, Caen, France, 22-25 Juin 2015.*

# Bibliographie


## A. Livres

1. Manning Christopher, Hinrich Schütze. (1999). *Foundations of Statistical Natural Language Processing*. ISBN 0-262-13360-l, Second printing, 1999 Massachusetts Institute of Technology.
2. Philip Koehn, Franz Josef Och, Daniel Marcu. (2003). *Statistical phrase-based translation*. In Proceedings of HLT-NAACL, 2003.
3. Jurafsky Daniel, James H. Martin. (2008). *Speech and Language Processing : An introduction to natural language processing*. Computational Linguistics and Speech Recognition, 2008.

## B. Articles (non exhaustifs)

1. Ehrmann, M. (2008). Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguısation. (Thèse de doctorat). Paris Diderot University.
2. Koehn, P. et Knowles, R. (2017). Six challenges for neural machine translation. arXivpreprint arXiv :1706.03872.
3. Mangeot, M. et Sadat, F. (2014). TALN-RECITAL 2014 workshop TALAf 2014 : Traitement automatique des langues africaines (TALAf 2014 : African language processing). Dans TALN-RECITAL 2014 Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014 : African Language Processing).

# Questions et Réponses

Merci de votre attention !