Pre-GDR TAL - ”Multilinguisme, multiplicité des langues”

Laurent Besacier, Massih-Reza Amini, Loïc Barrault, Delphine Bernhard, Olivier Kraif, Emmanuel Morin, Kamel Smaili, François Yvon

Juin 2018

1 Introduction

Ce chapitre présente un état des lieux des travaux en traitement automatique du langage naturel, écrit et oral, autour du thème *multilinguisme et multiplicité des langues*. De telles recherches s'appuient sur la disponibilité de ressources langagières qui couvrent de plus en plus de langues du monde et sur des ressources massivement parallèles qui nourrissent des approches véritablement multilingues en traduction automatique (Koehn, 2005), en analyse syntaxique (Nivre et al., 2016), en reconnaissance automatique de la parole (Schultz et al., 2013; Schultz and Schlippe, 2014), en désambiguisation lexicale (Navigli and Ponzetto, 2010; Sérasset, 2014), ou encore en dialectologie computationelle (Christodoulopoulos and Steedman, 2015), etc. Les aspects liés aux ressources sont cependant abordés dans un autre chapitre de ce document

Ce chapitre est donc structuré selon les thèmes suivants. La section 2 concerne la traduction automatique de l'écrit et de l'oral qui est actuellement confrontée à un changement de paradigme important depuis l'apparition de modèles neuronaux profonds. La section 3 est un panorama des approches d'alignement, de transfert et de similarité entre langues. Elle se consacre aussi aux récentes tentatives qui consistent à apprendre des modèles véritablement multilingues (un seul modèle pouvant être appliqué à plusieurs langues). La section 4 concerne le traitement automatique des langues peu dotées, des dialectes, et des langues régionales qui a émergé il y a une douzaine d'années et s'est fortement développé sous l'impulsion de projets d'envergure (par exemple projets financés par la DARPA[1]). La section 5 aborde d'autres thèmes (TAL et apprentissage des langues, TAL et dialectologie, problèmes d'alternance de code). Enfin, la section 6 présente quelques éléments de programmatique pour cet axe.

Le traitement de l'information multilingue et la communication multilingue répondent par ailleurs à de multiples enjeux du monde socio-économique : accès à l'information et interaction pour tous et dans toutes les langues, sécurité et surveillance, fouille dans des données massives, documentation et revitalisation des langues menacées, etc.

[1] Voir projets LORELEI, BABEL, etc.

2 Traduction automatique de l'écrit et de l'oral

La traduction automatique reste une des applications emblématiques du TAL (et, depuis quelques années, une application de choix pour l'apprentissage automatique), et probablement même de l'IA, tant la production d'une bonne traduction peut mobiliser, dans certains cas, une somme de connaissances (sur les langues, sur les cultures, sur les civilisations, etc.) ainsi qu'une créativité et une sensibilité qui sont souvent considérées comme l'apanage des plus hautes formes d'intelligence humaine.

Des avancées tangibles ont été réalisées ces dernières années (Bahdanau et al., 2014; Vaswani et al., 2017), qui ont permis de faire progresser l'état de l'art et rendre la TA visible et utile pour une large gamme d'applications. Les modèles les plus courants sont composés d'un encodeur bi-directionnel utilisant des unités récurrentes (GRU ou LSTM), associé à un décodeur (lui aussi composé de GRU ou LSTM) et équipé d'un mécanisme d'attention permettant de focaliser sur une partie spécifique de l'entrée lors de la production d'un mot en sortie (Bahdanau et al., 2014). Une séquence de symboles (dont la représentation est vectorielle) est calculée itérativement en appliquant à chaque pas de temps la même fonction apprise pour calculer le nouvel état du réseau à partir de l'état précédent et du dernier symbole produit dans la séquence en cours de production. Plus récemment, les modèles sans unités récurrentes sont apparus tels que le modèle *Transformer* (Vaswani et al., 2017). Grâce à ces progrès, les outils de traduction automatique en ligne[2] sont désormais largement utilisés par le grand public (par exemple traduction de descriptions et d'avis sur des sites de voyages) ou par les traducteurs professionnels (fourniture d'une ou plusieurs pre-traductions dans une interface de traduction assistée - un paragraphe est consacré plus spécifiquement à ce problème plus loin).

Ces avancées portent en particulier sur le design et l'estimation de grands modèles statistiques et neuronaux à partir de corpus parallèles (Koehn, 2005), multi-parallèles, voire monolingues (He et al., 2016; Wu et al., 2017; Conneau et al., 2017b; Artetxe et al., 2018; Lample et al., 2018a,b). Il reste cependant un long chemin à accomplir et de nombreux obstacles à surmonter pour atteindre l'objectif d'une traduction automatique performante, transparente et fiable, pour tous les couples de langues et contextes de traduction.

Addiction aux données. Un premier axe de travail concerne l'optimisation des données : les méthodes empiriques (probabilistes, neuronales) de TA demandent des corpus parallèles, qui sont notoirement insuffisants en taille, souvent trop spécialisés, et insuffisamment divers puisqu'il n'en existe que pour un très petit petit nombre de styles, registres, domaines, et paires de langues. Pour aborder ce problème, il existe un foisonnement de travaux en cours autour de la traduction multilingue (Johnson et al., 2016b), de l'adaption au domaine (Chu et al., 2017) et plus généralement au registre via l'apprentissage sans données parallèles (Artetxe et al., 2018; Conneau et al., 2017b; He et al., 2016; Lample et al., 2018a,b; Wu et al., 2017). Cependant, pour la traduction d'échanges informels (SMS, chats, tweets) ou la traduction de parole (Cettolo et al., 2017), les avancées sont plus lentes, faute en particulier de disposer de données appropriées, voire d'une définition claire de la tâche (doit on traduire les émoticons ? les

[2] `https://translate.google.com` ou `https://www.deepl.com`

hésitations ? Les faux départs ? Les reprises ? Les erreurs de grammaire volontaires ou involontaires ? L'intonation ? etc.).

Utilisation de connaissances expertes. Une manière alternative d'aborder ce problème, plus typique de travaux en TAL, est d'injecter des connaissances supplémentaires pour assister le processus de calcul d'une traduction automatique. Cette approche a fait dans une certaine mesure ses preuves pour la génération précédente de modèles statistiques (Costa-juss et al., 2016), et permet d'esquisser une réponse à l'aporie, souvent entendue chez les praticiens de l'apprentissage automatique, que la TA de haute qualité passerait par l'utilisation de moins, plutôt que plus, de connaissances expertes. L'introduction de connaissances expertes, symboliques, sur la langue et/ou le contexte en traduction automatique neuronale est une question très ouverte - la floraison de travaux récents sur l'utilisation de syntaxe en langue source (Aharoni and Goldberg, 2017; Bastings et al., 2017; Li et al., 2017; Tran and Bisk, 2018) est un signe d'un mouvement dans cette direction mais beaucoup reste encore à faire pour intégrer des connaissances sémantiques, ou pour prendre en compte la structure du discours. Ce qui est dit de l'utilisation d'une analyse profonde en source vaut aussi naturellement pour la langue cible, demandant pour implémenter de telles stratégies des analyses robustes des traductions produites, et la capacité de traduire au-delà de simples phrases isolées. Pour ce qui concerne l'introduction de connaissances contextuelles, la modélisation neuronale a ouvert de nouvelles portes et des perspectives excitantes: utilisation de contextes longs (à l'échelle du paragraphe ou du document, pour traduire les phénomènes au delà de la phrase : co-références Bawden et al. (2018), cohérence stylistique d'une partie du discours à l'autre) en source et en cible Tiedemann and Scherrer (2017); Jean et al. (2017), utilisation d'images ou d'illustrations comme une forme de supervision faible (*grounding*) de la traduction (Elliott et al., 2017a).

Traduction assistée. Un autre scénario d'usage de la TA dans lequel une collaboration avec des connaissances expertes est souhaitable est celui de l'outillage du traducteur et la traduction assistée par ordinateur (TAO). Une première question est de fournir au traducteur un meilleur contrôle des calculs de la TA, en lui permettant d'imposer a priori (ou a posteriori) des choix de traduction Hokamp and Liu (2017); **?** ou toute autre contrainte qui pourrait sembler utile (par exemple sur la structure des phrases dans un cadre de traduction à destination de certains publics) ; une seconde est d'assurer une plus grande transparence des choix et une plus grande systématicité de la TA, un problème qui s'applique bien plus largement pour tous les systèmes à base de réseaux profonds Belinkov et al. (2017a,b); Belinkov and Glass (2017); une troisième enfin consistera à intégrer le caractère nécessairement évolutif et personnalisé de la TA "outil du traducteur": l'apprentissage adaptatif en ligne, l'apprentissage sur le très long terme de la TA posent encore des défis considérables qu'il faudra savoir relever. Travailler sur la tâche de révision automatique (Bojar et al., 2017), une activité qui se développe dans l'industrie, reste d'actualité, et ce d'autant plus que les TA neuronales présentent un biais envers la fluidité des sorties, parfois au détriment de leur fidélité (Koehn and Knowles, 2017); développer de nouvelles mesures de confiance, voire de nouvelles métriques plus diagnostiques servira les mêmes buts. De manière plus fondamentale, les applications de TAO invitent à repenser à la fois les systèmes de traitement automatique aussi bien que les interfaces (par le clavier, par la voix, par le mouvement) par lesquelles le traducteur appréhende et manipule les textes source et cibles,

en accédant simultanément aux ressources (terminologies, mémoires) indispensables à son travail - quelques travaux **(REF)** commencent à aborder ces questions, qui engagent à une plus forte collaboration avec le domaine des interfaces homme-machine.

Gap sémantique. Par ailleurs, le *mur du sens* (ou *gap sémantique*) est aussi un enjeu majeur de la traduction automatique. En effet, les mots sémantiquement ambigus représentent un défi : pour produire une phrase correcte dans la langue cible, le système doit décider quelle signification est exacte dans le contexte source donné. Cependant, les modèles actuels emploient des techniques robustes issues de l'apprentissage automatique qui, tout en permettant de traiter de grands volumes de données, traduisent d'un langue à l'autre sans véritablement prendre en compte le sens de la phrase source. Les méthodes issues de la sémantique computationnelle peinent, quant à elles, à être utilisées dans les systèmes de traduction automatique et au final les interactions entre ces deux disciplines restent limitées. Une meilleure intégration des approches de désambiguisation léxicale et de clarification du sens pour la traduction automatique (Gonzales et al., 2017) est sans aucun doute un axe de recherche prometteur. Des travaux explorent également l'utilisation de données multimodales avec pour objectif, non seulement l'amélioration des résultats, mais également l'établissement de liens entre les concepts exprimés dans différentes modalités. La campagne d'évaluation internationale sur la traduction automatique multimodale tente de faire avancer l'état de l'art dans ce sens (Specia et al., 2016; Elliott et al., 2017b). Les systèmes multimodaux, quasiment tous fondés sur les réseaux de neurones, visent à relier et intégrer les informations textuelles et visuelles afin d'améliorer les performances sur un corpus multilingue composé de descriptions d'images (Elliott et al., 2016). Enfin, cet aspect est aussi important pour évaluer correctement les systèmes de traduction automatique (mais pas seulement) à partir de similarités sémantiques translingues robustes.

Traduction de la parole. Les systèmes de traduction de la parole actuels sont constitués de l'enchaînement de deux modules[3] : le premier effectue la reconnaissance de la parole (parole vers texte en langue source) et le second traduit automatiquement les sorties (sous forme de chaînes ou de graphes de mots en langue source) du premier module vers un texte en langue cible. Récemment, de premiers travaux sur la traduction directe *end-to-end* sans passer par les transcriptions en langue source ont été proposés (Berard et al., 2016; Weiss et al., 2017; Bansal et al., 2017; Bérard et al., 2018). Ceci pose la question générale suivante: pour traduire à partir d'un énoncé oral dans une langue source, a-t-on besoin de passer par une représentation intermédiaire symbolique (transcription orthographique ou phonétique) de cet énoncé ? La traduction de parole directe est donc un sujet très amont, dont les retombées pourraient être importantes pour la traduction automatique de dialectes oraux (données exploitables sans transcription de la langue source) et pour la traduction de parole en général (le temps de décodage d'un système de traduction *end-to-end* pourrait être moins important que pour un système qui enchaîne les étapes de transcription et de traduction, l'empreinte mémoire du modèle plus petite, etc.). L'intérêt croissant pour ce problème nouveau est d'ailleurs illustré par l'organisation d'une campagne d'évaluation lors de la conférence IWSLT 2018 où une condition *end-to-end* a été introduite (un seul modèle pour la tâche

[3]Si l'on fait abstraction de la synthèse vocale (TTS) vers la langue cible.

de traduction de parole).[4]

3 Alignement, transfert, similarités entre langues

Alignement. Les modèles de traduction automatique offrent aussi un cadre pour modéliser la mise en correspondance entre langues au niveau le plus fin: celui des mots, voire des unités sous-phrastiques. A la différence de la traduction automatique, l'alignement est une tâche non-supervisée et probablement mal définie, ce qui en renforce la difficulté. Calculer les alignements dans des textes bilingues a des applications variées: extraction de lexiques (Morin and Daille, 2012), recherche d'information interlingue (Nie, 2010), documentation automatique de langues peu dotées (Adda et al., 2016; Anastasopoulos and Chiang, 2017; Godard et al., 2016a), apprentissage des langues, etc. L'alignement s'applique sur des collections de textes traduits (c'est-à-dire des corpus alignés) et cherche à apparier dans les deux langues des unités textuelles de grain moindre que le texte : paragraphe, séquence phrastique, séquence intraphrastique, sans caractérisation linguistique (Véronis, 2000). Désormais, les travaux en alignement exploitent d'autres types de corpus que les corpus alignés notamment des corpus comparables (Zweigenbaum and Benoît, 2006), relevant de deux langues mais sans être en rapport de traduction, et des corpus multimodaux, relevant de deux modalités comme un texte écrit et sa retranscription de l'oral (Robert-Ribes et al., 1997) ou un texte image et sa transcription (Toselli et al., 2011).

En ce qui concerne les corpus comparables, les segments alignés sont le plus souvent des mots, des termes simples et des termes complexes. L'alignement à partir de corpus comparables s'est principalement concentré sur les mots simples en domaine de langue général (Fung, 1998; Rapp, 1999; Gaussier et al., 2004; Mikolov et al., 2013) et sur les termes simples (Chiao and Zweigenbaum, 2002; Morin et al., 2007; Bouamor et al., 2013) et complexes (Morin and Daille, 2012) en domaine de spécialité en s'appuyant sur des méthodes distributionnelles ou distribuées. Les travaux les plus récents du domaine s'inscrivent dans le mouvance des approches basées sur les réseaux de neurones (Mikolov et al., 2013; Jakubina and Langlais, 2017; Hazem and Morin, 2017) pour les mots et termes simples.

Les lexiques bilingues extraits de corpus comparables représentent des données précieuses car ils permettent d'accéder au vocabulaire originel d'une domaine spécialisé ou technique sans biais induit par un mécanisme de traduction. Ces lexiques intéressent notamment les traducteurs professionnels (Delpech, 2014) lors de l'étape de révision d'un texte à traduire en particulier lorsque les termes sont illustrés par des contextes permettant d'en appréhender leurs usages.

Un des défis à venir concerne l'alignement de termes de longueurs différentes (par exemple l'alignement d'un terme simple avec un terme complexe) à partir de corpus comparables en domaine de spécialité. De manière concomitante, l'abondance des ressources disponibles (corpus comme lexiques) est un autre défi pour sélectionner les données les plus pertinentes à ajouter aux modèles existants.

Un autre défi consiste à aligner des corpus multimodaux (parole en source et texte

[4] voir `https://sites.google.com/site/iwsltevaluation2018/Lectures-task`.

en cible, par exemple). Des travaux pionniers sur ce domaine ont été réalisés et semblent particulièrement intéressants à développer (Anastasopoulos and Chiang, 2017; Duong et al., 2016b). Les premières tentatives d'alignement direct entre signal en langue source et texte en langue cible sont les travaux de Duong et al. (2016a). Les mêmes auteurs (Anastasopoulos et al., 2016) proposent également d'utiliser conjointement les modèles IBM de traduction (IBM Model 2) et l'alignement dynamique de signaux (DTW) pour aligner de la parole source et du texte cible. Enfin, le projet ANR Franco-Allemand BULB (Adda et al., 2016) s'intéresse à l'extraction non supervisée de lexiques à partir de corpus de parole parallèles (Godard et al., 2016b).

Transfert. La plupart des systèmes de traitement automatique du langage naturel sont fondés sur des modèles entraînés sur de grands corpus et correspondant à une langue cible donnée. Récemment, des approches par transfert ou projection, consistant à projeter rapidement des modèles d'un langue à une autre, sont apparues. Le point commun de ces approches est de trouver et d'explorer des mécanismes non (ou très peu) coûteux pour exploiter des ressources linguistiques annotées déjà disponibles pour certaines langues et des corpus parallèles ou comparables pour produire de nouvelles ressources annotées pour d'autres langues plus faiblement dotées. La transfert consiste donc à identifier des équivalences morpho-syntaxiques (Yarowsky et al., 2001; Wisniewski et al., 2014; Zennaki et al., 2016), syntaxiques Hwa et al. (2005); Tiedemann (2014); Lacroix et al. (2016); Aufrant et al. (2016) ou sémantiques (Padó and Lapata, 2009; Jabaian et al., 2013) à partir de corpus de textes parallèles ou comparables. De telles approches semblent particulièrement intéressantes pour construire des systèmes performants dans des scénarios à faibles ressources où la quantité de données d'apprentissage pour une langue ou un groupe de langues données est limitée.

Plus récemment, avec la montée en puissance de l'apprentissage de modèles de bout-en-bout (*end-to-end*), on s'est rapproché du Graal consistant à construire des systèmes véritablement multilingues (un seul système pour toutes les langues). Par exemple, les approches neuronales (notamment encodeur-décodeur) permettent aisément de modéliser plusieurs langues dans un seul système, pour autant qu'on puisse définir une représentation multilingue des entrées et sorties (par ex. à base de caractères, ou d'unités sous-lexicales (Sennrich et al., 2015)). Différentes approches sont actuellement considérées. Elles partagent l'objectif de mutualiser certaines parties du modèle neuronal afin que les multiples langues enrichissent le modèle et le rendent plus robuste (Ha et al., 2016; Johnson et al., 2016a; Gu et al., 2018). L'un des principal enjeu est donc de définir un espace de représentation commun à toute les langues. Cet espace pourrait être assimilé à un *interlingua*, permettant d'obtenir des représentations abstraites indépendantes de la langue.

Similarités Les méthodes de transfert, mentionnées dans le paragraphe précédent, s'appuient sur des représentations de mots ou phrases dans un espace multilingue apprises à partir de corpus parallèles (Mikolov et al., 2013), de dictionnaires **(REF)** ou sans véritable ressource disponible au préalable (Conneau et al., 2017a; Conneau and Kiela, 2018). Ces méthodes se sont largement développées ces dernières années grâce notamment au déploiement massif des modèles neuronaux. Cet intérêt s'explique notamment par la capacité de ces modèles à apprendre une représentation des données dans un espace de plongement de façon totalement générique, ce qui ouvre le champ au développement de différentes techniques d'appariement au niveau mot ou même docu-

ment. Plusieurs articles parus récemment dans les conférences et les revues du domaine (Mrksic et al., 2017; Rajendran et al., 2016; Vulic, 2017; Zhang et al., 2017) témoignent de cet attrait. De tels travaux trouvent aussi des applications dans d'autres domaines tels que la détection de plagiat translingue (Schwab et al., 2017) et la recherche d'information multilingue (Balikas et al., 2018).

Le transfert entre langues est favorisé par les proximités entre langues, et est plus simple lorsqu'il s'agit de langues d'un même famille linguistiques, voire de variantes d'une même langue souche . L'exploitation délibérée de proximités typologiques Naseem et al. (2010, 2012) peut se faire de multiples manières, par exemple via des biais sur les a priori, ou encore en utilisant les ressources telles que le WALS **?**.

[LB: Compléter + Défis?]

Mentionner récent group de travail sur *Special Issue on NLP for Similar Languages, Varieties and Dialects* avec numéro spécial à venir sur le sujet.[5]

4 Langues peu dotées, dialectes, langues régionales

Les recherches en TAL pour les langues peu dotées s'appuient sur la disponibilité de ressources langagières qui couvrent de plus en plus de langues du monde. Ceci est illustré par l'évolution des ressources du *LRE Map*[6] analysée par J. Mariani au cours de l'atelier LREC/CCURL 2018[7] qui met en évidence une forte progression du nombre de ressources proposées pour un nombre croissant de langues du monde.[8] En Europe, on observe également une forte croissance de la disponibilité de ressources pour des langues régionales européennes.[9]

Définition. Les langues que l'on nomme "peu dotées" sont des langues pour lesquelles il n'existe que peu ou pas de ressources et d'outils de traitement automatique. Cette notion couvre en réalité un large spectre de réalités linguistiques, allant des langues officielles disposant de peu de locuteurs et/ou peu de moyens aux dialectes et langues régionales n'ayant aucune reconnaissance officielle **+ sociolectes??**. À cela s'ajoutent les états anciens des langues, qui constituent des objets d'étude en linguistique diachronique et humanités numériques, ainsi que les variantes régionalement marquées de langues par ailleurs bien dotées. Le potentiel de l'application du traitement automatique à ces langues reste largement sous-exploité, notamment dans le but de les documenter, voire, pour certaines, de les revitaliser. Le rôle du TAL dans la documentation des langues est avant tout de faciliter le travail des linguistes de terrain et sociolinguistes, en assistant les tâches de collecte et d'analyse de corpus (Nguyen et al., 2016; Blachon et al., 2016; Adda et al., 2016) ou en permettant de nouvelles

[5] `https://sites.google.com/view/nledialects`

[6] Initié par ELRA lors de la conférence LREC 2010, LRE Map est un mécanisme destiné à recenser l'utilisation et la création de ressources langagières en recueillant des informations sur les ressources existantes et nouvellement créées au cours du processus de soumission d'articles. Près de 8000 formulaires de ressources ont été remplis depuis 2010 au cours de 16 conférences telles que LREC, COLING, IJCNLP, LTC, etc.

[7] `http://www.ilc.cnr.it/ccurl2018/`

[8] Europe (+300%) ; Afrique (+150%) ; Amérique du nord (+1200%) ; Amérique du sud (+350%) ; Asie (+450%)

[9] Une centaine en 2010 et plus de 250 en 2016, couvrant 32 langues européennes régionales au total

avancées en typologie et phylogénétique des langues (Murawaki, 2015; Asgari and Schütze, 2017; Malaviya et al., 2017). Pour ce qui est de la revitalisation, on constate actuellement un regain d'intérêt pour certaines langues en perte de locuteurs via l'utilisation des réseaux sociaux[10].

Variabilité. D'une manière générale, les langues peu dotées se caractérisent par une forte variation à l'oral et à l'écrit (lorsqu'elles disposent d'un système d'écriture), ce qui constitue un défi pour les outils de TAL, avec en plus la forte contrainte du manque de données disponibles (*small data*). Ces défis favoriseront dans les années à venir les travaux visant à mettre au point des méthodes extrêmement robustes et peu sensibles au manque de données, qui viendront compléter les travaux actuels se focalisant sur le traitement de grandes quantités de données pour les langues bien dotées (*big data*). Le double défi de la variation et du manque de données constitue également une formidable opportunité pour les outils et ressources existants car il conviendra de mieux exploiter les proximités entre langues pour faciliter leur réutilisation, grâce notamment à des techniques de transfert d'annotations déjà évoquées dans la section précédente (Yarowsky and Ngai, 2001; Guo et al., 2015). Sur ce point, le critère du choix des langues sources (doit on s'appuyer sur des langues proches *bien dotées* de la même famille que la langue cible ?) reste encore peu abordé et fera sans aucun doute l'objet de recherches intéressantes dans le futur. Des travaux actuels exploitant des traits issus de bases typologiques (telles que WALS[11]) vont actuellement dans ce sens, mais il reste beaucoup à faire dans la recherche de généralisations et de traits partagés entre multiples langues afin d'induire des connaissances relativement poussées sur un langue peu dotée donnée.

Collecte. Un autre défi des langues peu dotées est la collecte de ressources. Des approches innovantes par crowdsourcing éthique pour l'écrit (Millour and Fort, 2018) ou l'oral (voir par example l'initiative *Common Voice*[12] de la fondation Mozilla) sont récemment apparues. L'utilisation de jeux sérieux[13] (Guillaume et al., 2016) pour la collecte des données et d'annotations en TAL ainsi que l'utilisation d'applications mobiles pour la collecte de ressources orales[14] (Blachon et al., 2016) sont des approches qui devraient se développer dans le futur mais un autre défi reste de réussir à impliquer activement les principaux intéressés qui sont les locuteurs des langues peu dotées ciblées.

Approches zero-shot.

On ne peut pas ne pas mentionner dans cette section les approches dites *zero ressource* (ou *zero shot*) qui sont apparues récemment en TAL. Le but de ces approches non supervisées est de résoudre une tâche d'apprentissage sans avoir reçu d'exemples annotés correspondant à ladite tâche. Par exemple, dans le domaine du traitement automatique de la parole, des recherches ont été initiées via le défi *Zero Resource*

[10] **Les études d'envergure sur l'utilisation des langues peu dotées, et notamment les langues régionales, sur les réseaux sociaux manquent encore. Cela étant, ce type de données a déjà été exploité dans des travaux de TAL pour la constitution de corpus ou des études spécifiques (Frey et al., 2015; Burghardt et al., 2016)**

[11] http://wals.info

[12] https://blog.mozilla.org/blog/2018/06/07/parlez-vous-deutsch-rhagor-o-leisiau-i-common-voice/

[13] http://anawiki.essex.ac.uk/dali/games4nlp/

[14] https://lig-aikuma.imag.fr

Speech Challenge[15] (Dunbar et al., 2017) dont l'objectif ultime est de construire un système qui apprend à dialoguer oralement dans une langue inconnue, à partir de zéro, en utilisant uniquement les informations disponibles pour un enfant en bas âge (pas de données transcrites, juste de la parole brute plus éventuellement une information visuelle ou une rétroaction humaine, etc.). Le fait que les enfants de 4 ans apprennent une langue sans une supervision correspondant à une transcription exacte de ce qui est dit, montre que cet objectif est théoriquement réalisable. Au delà de la question fondamentale de l'acquisition du langage (comment un système peut-il acquérir le langage de façon autonome ?), résoudre ce type de défi permettrait de s'affranchir de l'addiction aux données des systèmes de traitement automatique de la parole actuels. En effet, la plupart des langues du monde n'ont pas de ressources textuelles ou même une orthographe fiable. Des systèmes construits sans ressources expertes pourraient servir des millions d'utilisateurs de ces langues dites non écrites. Ces technologies pourraient aussi aider les linguistes de terrain à analyser (semi-)automatiquement et à annoter les enregistrements audio de ces langues en danger avec des unités linguistiques découvertes automatiquement (phonèmes, lexique, lexique, grammaire) (Adda et al., 2016).

Défi technologique et sociétal. Afin d'encourager l'utilisation des langues orales via de nouveaux médias tels que les réseaux sociaux, il sera nécessaire de développer des méthodes adaptées pour les claviers prédictifs (ou la correction orthographique et la complétion automatique), capables de gérer la forte variation graphique et l'alternance codique (sur le sujet de l'aternance codique, voir par exemple (El-Haj et al., 2018) ou (Mendels et al., 2018)). Une autre direction de recherche prometteuse concerne le développement d'outils et de ressources pour l'apprentissage des langues en danger assisté par ordinateur (Maxwell and Bills, 2017; Katinskaia et al., 2017). Toutes ces applications du TAL pourraient avoir un fort impact sociétal. On peut aussi se poser la question de l'impact qu'auront ces technologies (et leur différents degrés d'évolution selon les langues) sur les langues du monde et si elles permettront leur revitalisation. Par exemple, Ostler (2010) suggère que l'anglais sera la dernière *lingua franca* et que ce qui lui succèdera ne sera pas une autre langue unique. Il fait plutôt l'hypothèse d'un retour à l'état de Babel grâce aux progrès de la traduction informatique où "chacun parlerait et écrirait dans la langue de son choix".

5 Autres aspects liés au multilinguisme en TAL [LB: trouver une autre formulation]

TAL et apprentissage des langues étrangères

[LB: à compléter + references à intégrer dans le bibtex]

[LB: @olivier: pourras tu integrer les bibtex de toutes tes references stp ?]

Pour les TICE et l'enseignement/apprentissage des langues, les technologies du TAL connaissent des applications variées, dans des domaines généralement identifiés comme ICALL (Intelligent Computer Assisted Language Learning) ou NLP for CALL.

[15] `http://zerospeech.com`

(**?**) relèvent les axes de développement suivants, pour lesquels on peut citer quelques applications effectives dans des curriculums de formation :

- Génération de ressources (contenus de référence pour la didactique) à partir de corpus : fouille de corpus, corpus bilingues, interface dictionnaire/corpus
- Aide à la lecture : enrichissement d'annotations permettant d'accéder à des dictionnaires, grammaires, conjugueurs, etc. (p.ex. Spanish for Business Professional, Hagen, 1999)
- Génération d'activités : génération d'exercices (exercice lacunaire, exercice de prononciation, etc.) (Alfalex, Selva, Verlinde, Binon, 2004)
- Détection d'erreur et génération de feed-back, évaluation automatique : en général pour des réponses courtes et contraintes par le contexte de l'activité (p.ex. Etutor, Heift, 2005 ; Tagarella, Amaral-Meurers, 2011)
- Aide à la rédaction **(REF)**
- Sélection automatique de textes **(REF)**
- Adaptation de l'environnement en fonction du modèle d'apprenant (Tagarella, Amaral-Meurers, 2011)

On constate dans les faits que ces technologies sont encore assez peu utilisées par les enseignants et les apprenants - la fiabilité des systèmes n'étant jamais de 100%, et le contexte didactique tolérant mal les erreurs d'analyse et les textes erronés aux plans lexical ou morphosyntaxique. Pourtant des pistes prometteuses ont encore été assez peu explorées, et il existe de grandes marges de progression : par exemple avec le traitement des corpus d'apprenants (Meurer, 2015), les systèmes d'aide à l'évaluation (Tack et al., 2016 ; Bestgen Y. 2016) ou à l'annotation pour les enseignants (Hamel et al., 2016), ou encore les systèmes d'aide à la fouille de corpus pour les pédagogies basées sur le Data Driven Learning (Johns, 2002 ; Chambers, 2010 ; Yang, 2017 ; Boulton et Cobb, 2017).

mentionner le défi CAP "my tailor is rich!" `http://cap2018.litislab.fr/competition.html` Le but de cette compétition est de réaliser, par apprentissage, un système permettant de prédire le niveau de compétence d'un apprenant, à partir d'une de ces productions écrites comprenant entre 20 et 300 mots et d'un ensemble de caractéristiques calculées à partir de ce texte.

TAL et Dialectologie

[LB: à qui demander ?]

Dialecto computationnelle et utilisation de bases typologiques pour le TAL

Parler des ref ci-dessous

Found in Translation: Reconstructing Phylogenetic Language Trees from Translations `https://arxiv.org/abs/1704.07146`

Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages `http://aclweb.org/anthology/D17-1011`

TAL et alternance de code

Dans certaines communautés, la conversation obéit à un phénomène particulier qu'on appelle la diglossie où deux langues coexistent, avec une langue formelle et une plus familière. C'est le cas notamment dans le monde arabe où la langue officielle est l'arabe standard, mais la population préfère parler plutôt sa langue maternelle qui est le dialecte.

Ce problème pose plusieurs problèmes en traitement automatique des langues puisque le dialecte est une langue vernaculaire qui ne dispose pas de ressources de type : corpus, dictionnaires, ontologies, outils, etc.

Par ailleurs, l'avènement des réseaux sociaux a permis d'alimenter le web de données dialectales importantes avec toutes les faiblesses grammaticales d'écriture que l'on peut imaginer. Ajouter à cela, un nouveau phénomène s'est greffé à la diglossie : le code-switching vers les langues occidentales. En effet, dans une conversation, les interlocuteurs en plus de l'arabe standard et du dialecte, ils basculent souvent vers les français, l'anglais, etc.

Ce phénomène pose de réels défis scientifiques. En reconnaissance de la parole, par exemple si le système est appris sur l'arabe standard, il échouera complètement lorsqu'il doit traiter des mots en dialecte ou en français. Ainsi, dans l'expérience menée au Loria, le WER sur un texte en arabe standard est de 14%, alors qu'il chute dramatiquement à 89% lorsqu'on essaie de reconnaître le dialecte mélangé au français. Ce phénomène de code switching n'est pas lié seulement à la communauté arabe, mais on le retrouve dans la communauté turque, indienne et autres.

M. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaili, "Development of the arabic loria automatic speech recognition system (ALASR) and its evaluation for Algerian dialect," in Third International Conference On Arabic Computational Linguistics, Dubai, 2017

Y. Yeong and T. Tan, "Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information," in INTERSPEECH, 15th Annual Conference of the International Speech Communication Association, 2014

A. Alhazmi, "Linguistic aspects of arabic-english code switching on facebook and radio in Australia," International Journal of Applied Linguistics and English Literature, vol. 5, no. 3, 2015.

Liens avec l'axe données

à clarifier quel ax aborde cette partie...

Standards généraux, cf Universal dependencies Interoperabilité des ressources multilingues (bases lexicales multilingues?) framenet qui sert de cadre à d'autres projets sur d'autres langues ? EPL: PARSEME

6 Eléments de programmatique

Au vu de cet état des lieux et des forces en présence en France, nous pensons qu'il est important de nous appuyer, dans le cadre du GDR, sur deux thèmes déjà forts en France et de développer deux thèmes plutôt en émergence qui présentent néanmoins un potentiel important dans le futur.

Les thèmes forts à développer nous semblent être

- La traduction automatique (et la TAO) ; thème sur lequel des laboratoires sont bien positionnés au niveau international (LIMSI, LIG, LIUM, LINA, LORIA) ; on peut aussi mentionner la présence ou l'arrivée récente en France de gros acteurs industriels qui abordent ce thème (Facebook, Naver, Google, Systran),

- Le traitement des langues peu dotées, des langues régionales et des dialectes ; thème sur lequel des acteurs académiques historiques sont présents en France (LIG, USTRA, LINA, LIMSI, LORIA).[16]

Les thèmes émergeants à développer nous semblent être

- Le TAL pour l'apprentissage des langues ; avec les progrès dans différents domaines du TAL et du TALP, certaines technologies semblent désormais matures pour aborder des tâches telles que l'analyse de productions textuelles et orales d'apprenants d'une langue seconde, pouvant contribuer, par exemple, à la revitalisation de langues en danger de disparition ; intérêts au LIDILEM, LIMSI, LIG, LORIA

- Le TAL et la dialectologie qui, dans un sens consiste à utiliser des bases dialectologiques massives (par exemple WALS) pour aider les systèmes de TAL et d'autre part contribuer, via les méthodes modernes de TAL, à la dialectologie (dialectologie computationnelle) ; intérêts et débuts de travaux au LIF, LIMSI, LIG, DDL.

précisier le lien avec GDR INSHS notamment sur les aspects dialecto et langues peu dotées

References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., Van de Velde, M., Yvon, F., and Zerbian, S. (2016). Breaking the unwritten language barrier: The Bulb project. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia.

Aharoni, R. and Goldberg, Y. (2017). Towards string-to-tree neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–140, Vancouver, Canada. Association for Computational Linguistics.

Anastasopoulos, A. and Chiang, D. (2017). A case study on using speech-to-translation alignments for language documentation. pages 170–178. Association for Computational Linguistics.

[16]Par exemple, Laurent Besacier (LIG) est *chair* d'un groupe d'intérêt spécial associé aux sociétés savantes ISCA (speech) et ELRA (language resources) sur les langues peu dotées: *Special Intersted Group for Under-resourced Languages (SIGUL)* - voir `http://elra.info/en/sig/sigul/`

Anastasopoulos, A., Chiang, D., and Duong, L. (2016). An Unsupervised Probability Model for Speech-to-Translation Alignment of Low-Resource Languages. *arXiv preprint arXiv:1609.08139*.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Asgari, E. and Schütze, H. (2017). Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Aufrant, L., Wisniewski, G., and Yvon, F. (2016). Zero-resource dependency parsing: Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 119–130, Osaka, Japan. The COLING 2016 Organizing Committee.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Balikas, G., Laclau, C., Redko, I., and Amini, M. (2018). Cross-lingual document retrieval using regularized wasserstein distance. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 398–410.

Bansal, S., Kamper, H., Lopez, A., and Goldwater, S. (2017). Towards speech-to-text translation without speech recognition. In *EACL (short papers)*, Valence (Spain).

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Simaan, K. (2017). Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. (2017a). What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Belinkov, Y. and Glass, J. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Advances in Neural Information Processing Systems*, pages 2438–2448.

Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2017b). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1–10.

Bérard, A., Besacier, L., Kocabiyikoglu, A. C., and Pietquin, O. (2018). End-to-End Automatic Speech Translation of Audiobooks. In *ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Alberta, Canada.

Berard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain.

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proceedings of SLTU (Spoken Language Technologies for Under-Resourced Languages)*, Yogyakarta, Indonesia.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *WMT*, pages 169–214. Association for Computational Linguistics.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2013). Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 759–764, Sofia, Bulgaria.

Burghardt, M., Granvogl, D., and Wolff, C. (2016). Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. In *Proceedings of LREC 2016*, pages 2029–2033, Portorož, Slovenia.

Cettolo, M., Federico, M., Bentivogli, L., Niehues, J., Stueker, S., Sudoh, K., Yoshino, C., and Federmann, C. (2017). Overview of the IWSLT 2017 Evaluation Campaign. In *International Workshop on Spoken Language Translation*, Tokyo, Japan.

Chiao, Y.-C. and Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Christodoulopoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391. Association for Computational Linguistics.

Conneau, A. and Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *CoRR*, abs/1803.05449.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017a). Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017b). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Costa-juss, M. R., Rapp, R., Lambert, P., Eberle, K., Banchs, R. E., and Babych, B. (2016). *Hybrid Approaches to Machine Translation*. Springer Publishing Company, Incorporated, 1st edition.

Delpech, E. M. (2014). *Comparable Corpora and Computer-assisted Translation*. John Wiley & Sons, Inc.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). THE ZERO RESOURCE SPEECH CHALLENGE 2017. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, Okinawa, Japan.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016a). An Attentional Model for Speech Translation Without Transcription. In *NAACL-HLT 2016*, Denver, Colorado, USA.

Duong, L., Anastasopoulos, A., Chiang, D., Bird, S., and Cohn, T. (2016b). An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT*, pages 949–959.

El-Haj, M., Rayson, P., and Aboelezz, M. (2018). Arabic Dialect Identification in the Context of Bivalency and Code-Switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017a). Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233. Association for Computational Linguistics.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017b). Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Frey, J.-C., Glaznieks, A., and Stemle, E. W. (2015). The DiDi Corpus of South Tyrolean CMC Data. In *Proceedings of the 2nd Workshop of the Natural Language Processing for Computer-Mediated Communication/Social Media*, pages 1–6.

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.-N., Löser, K., Rialland, A., and Yvon, F. (2016a). Preliminary Experiments on Unsupervised Word Discovery in Mboshi. In *Interspeech 2016*, San Francisco, California, USA.

Godard, P., Adda, G., Adda-Decker, M., Allauzen, A., Besacier, L., Bonneau-Maynard, H., Kouarata, G.-N., Löser, K., Rialland, A., and Yvon, F. (2016b). Preliminary Experiments on Unsupervised Word Discovery in Mboshi. In *Interspeech 2016*, San Francisco, California, USA.

Gonzales, A. R., Mascarell, L., and Sennrich, R. (2017). Improving word sense disambiguation in neural machine translation with sense embeddings. In *WMT*, pages 11–19. Association for Computational Linguistics.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. K. (2018). Universal neural machine translation for extremely low resource languages. *CoRR*, abs/1802.05368.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *COLING*, pages 3041–3052. ACL.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2015). Cross-lingual Dependency Parsing Based on Distributed Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Hazem, A. and Morin, E. (2017). Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP'17)*, pages 685–693, Taipei, Taiwan.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL'17, pages 1535–1546. Association for Computational Linguistics.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

Jabaian, B., Besacier, L., and Lefèvre, F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Trans. Audio, Speech & Language Processing*, 21(3):636–648.

Jakubina, L. and Langlais, P. (2017). Reranking translation candidates produced by several bilingual word similarity sources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 605–611, Valencia, Spain.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does Neural Machine Translation Benefit from Larger Context? *ArXiv e-prints*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016a). Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

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. (2016b). Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Katinskaia, A., Nouri, J., and Yangarber, R. (2017). Revita: a System for Language Learning and Supporting Endangered Languages. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa, Gothenburg, 22nd May 2017*. Linköping University Electronic Press.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *NMT@ACL*, pages 28–39. Association for Computational Linguistics.

Lacroix, O., Aufrant, L., Wisniewski, G., and Yvon, F. (2016). Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, pages 1058–1063, San Diego, California.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018b). Phrase-based & neural unsupervised machine translation. *CoRR*, abs/1804.07755.

Li, J., Xiong, D., Tu, Z., Zhu, M., Zhang, M., and Zhou, G. (2017). Modeling source syntax for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–697, Vancouver, Canada. Association for Computational Linguistics.

Malaviya, C., Neubig, G., and Littell, P. (2017). Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.

Maxwell, M. and Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91.

Mendels, G., Soto, V., Jaech, A., and Hirschberg, J. (2018). Collecting Code-Switched Data from Social Media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Millour, A. and Fort, K. (2018). Toward a lightweight solution for less-resourced languages: Creating a POS tagger for alsatian using voluntary crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Morin, E. and Daille, B. (2012). Revising the compositional method for terminology acquisition from comparable corpora. In *Proceedings of COLING 2012*, pages 1797–1810, Mumbai, India. The COLING 2012 Organizing Committee.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th*

Annual Meeting of the Association for Computational Linguistics (ACL'07), pages 664–671, Prague, Czech Republic.

Mrksic, N., Vulic, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gasic, M., Korhonen, A., and Young, S. J. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *TACL*, 5:309–324.

Murawaki, Y. (2015). Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334, Denver, Colorado. Association for Computational Linguistics.

Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 629–637, Stroudsburg, PA, USA. Association for Computational Linguistics.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1234–1244, Stroudsburg, PA, USA. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In Hajic, J., Carberry, S., and Clark, S., editors, *ACL*, pages 216–225. The Association for Computer Linguistics.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Nie, J.-Y. (2010). Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125.

Nivre, J., de Marneffe, M., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Ostler, N. (2010). *The Last Lingua Franca: English Until the Return of Babel*. Walker.

Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2016). Bridge correlational neural networks for multilingual multimodal representation learning. In *HLT-NAACL*, pages 171–181. The Association for Computational Linguistics.

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Robert-Ribes, J., Mukhtar, R. G., and Crc, A. C. S. (1997). Automatic generation of hyperlinks between audio and transcript. In *Proceedings of the 5th,European Conference on Speech Communication and Technology (Eurospeech '97)*, Rhodes, Greece.

Schultz, T. and Schlippe, T. (2014). Globalphone: Pronunciation dictionaries in 20 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 337–341.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8126–8130.

Schwab, D., Besacier, L., Ferrero, J., and Agnès, F. (2017). Using word embedding for cross-language plagiarism detection. In *EACL (2)*, pages 415–421. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sérasset, G. (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, pages – . To appear.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Tiedemann, J. (2014). Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92. Association for Computational Linguistics.

Toselli, A. H., Romero, V., and Vidal, E. (2011). Alignment between text images and their transcripts for handwritten documents. In Sporleder, C., van den Bosch, A., and Zervanou, K., editors, *Language Technology for Cultural Heritage*, pages 23–37, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tran, K. and Bisk, Y. (2018). Inducing grammars with and for neural machine translation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Vulic, I. (2017). Cross-lingual syntactically informed distributed word representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 408–414.

Véronis, J., editor (2000). *Parallel Text Processing: Alignment and use of translation corpora*. Kluwer Academic Publishers, Dordrecht.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly transcribe foreign speech. *arXiv preprint arXiv:1703.08581*.

Wisniewski, G., Pécheux, N., Gahbiche-Braham, S., and Yvon, F. (2014). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785, Doha, Qatar.

Wu, L., Xia, Y., Zhao, L., Tian, F., Qin, T., Lai, J., and Liu, T.-Y. (2017). Adversarial neural machine translation. *arXiv preprint arXiv:1704.06933*.

Yarowsky, D. and Ngai, G. (2001). Inducing multilingual POS taggers and NP brackets via robust projection across aligned corpora. In *Proceedings of NAACL*.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zennaki, O., Semmar, N., and Besacier, L. (2016). Inducing multilingual text analysis tools using bidirectional recurrent neural networks. In *COLING*, pages 450–460. ACL.

Zhang, M., Liu, Y., Luan, H., and Sun, M. (2017). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1959–1970.

Zweigenbaum, P. and Benoît, H. (2006). Faire se rencontrer les parallèles : Regards croisés sur l'acquisition lexicale monolingue et multilingue.