Étude de l'apprentissage par transfert de systèmes de traduction automatique neuronaux

Adrien BARDET[1] Fethi BOUGARES[1] Loïc BARRAULT[1]
(1) LIUM, adresse, 72000 Le Mans, France
prénom.nom@univ-lemans.fr

RÉSUMÉ

L'apprentissage par transfert est une solution au problème de l'apprentissage de systèmes de traduction automatique neuronaux pour des paires de langues peu dotées. Dans cet article, nous proposons une analyse de cette méthode. Nous souhaitons évaluer l'impact de la quantité de données et celui de la proximité des langues impliquées pour obtenir le meilleur transfert possible. Nous prenons en compte ces deux paramètres non seulement pour une tâche de traduction "classique" mais également lorsque les corpus de données font défaut. Enfin, il s'agit de proposer une approche où volume de données et proximité des langues sont combinées afin de ne plus avoir à trancher entre ces deux éléments.

ABSTRACT

Study on transfer learning in neural machine translation

Transfer learning is a solution to learn neural machine translation systems when dealing with low ressourced languages pairs. In this paper, We propose an analysis of transfer learning. We want to assess the correlation between data quantity and languages proximity to improve the transfer. We compare these parameters for transfer learning in a classical context and a low resource context. Finally, we want to propose an approach where data quantity and languages proximity are combined so that we do not have to choose between these two elements.

MOTS-CLÉS : apprentissage par transfert, traduction automatique neuronale, quantité de données, proximité des langues.

KEYWORDS: tranfer learning, neural machine translation, data quantity, languages proximity.

1 Introduction

Les performances des systèmes de traduction automatique neuronaux évoluent rapidement, cependant cela ne se verifie pas lorsque peu de données sont disponibles. L'apprentissage par transfert est une voie intéressante pour pallier à ce problème. Il consiste à entraîner un modèle sur une tâche bien dotée (modèle "parent"), puis le réutiliser pour l'apprentissage d'un modèle "enfant" (en remplacement d'une initialisation aléatoire). L'objectif est de capitaliser sur les représentations apprises par le système "parent". Ce transfert améliore généralement les résultats du système enfant par un transfert de connaissances apprises par le système parent (Zoph *et al.*, 2016).

Dans cet article, nous analysons différents critères ayant un impact sur l'apprentissage d'un système de traduction automatique neuronal par transfert. Dans notre cas, un premier système est entraîné sur une paire de langues puis réutilisé pour entraîner un second système pour la paire de langues ciblée.

Nous nous intéressons aux différents paramètres qui entrent en compte lorsque l'on emploie ce genre de technique, notamment les caractéristiques des données utilisées pour apprendre le système qui sert de base à notre transfert. En effet, plusieurs travaux portent sur les quantités de données utilisées ainsi que la proximité des langues mises en jeu, et les conclusions divergent (Kocmi & Bojar, 2018; Dabre *et al.*, 2017).

Nous cherchons donc à analyser les configurations de données pour l'apprentissage par transfert afin de déterminer les paramètres pertinents pour obtenir les meilleures performances.

2 Travaux Connexes

Certaines grandes avancées techniques ont permis au systèmes neuronaux de devenir l'approche la plus efficace pour la traduction automatique (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014).

Actuellement, les systèmes neuronaux nécessitent une grande taille de corpus d'entraînement afin d'obtenir de bonnes performances, ce qui, par définition, pose problème pour les paires de langue peu dotées. Les systèmes de traduction automatique statistique à base de segments sont alors une alternative pertinente (Koehn & Knowles, 2017).

Plusieurs approches de traduction automatique multilingue ont été utilisées pour traduire des textes dans des paires de langues peu dotées (Lakew *et al.*, 2018; Gu *et al.*, 2018; Johnson *et al.*, 2016). L'utilisation d'encodeurs et de décodeurs universels ont permis à Johnson *et al.* (2016) de concevoir un système apprenant en parallèle de nombreuses paires de langues et obtenant de meilleurs résultats, notamment pour les langues moins dotées. Des symboles spécifiques (ex : <2es>) sont alors utilisés afin de contrôler la langue en sortie du décodeur universel. Ce genre de modèle permet même de traduire dans des paires de langues non vues pendant l'entraînement (on parle alors de *zero-shot learning*). Cependant, les performances dans de tels cas restent relativement faibles. L'apprentissage par transfert peut combler ce manque en se basant sur un système parent (appris sur une grande quantité de données) qui sert de base pour apprendre un système enfant (Zoph *et al.*, 2016). Le transfert s'opère plus efficacement lorsque des langues sont partagées entre le parent et l'enfant. Dans cette direction, Dabre *et al.* (2017) met en avant l'importance de la proximité des langues pour que le transfert soit de meilleure qualité. Ces observations sont contredites dans Kocmi & Bojar (2018) où de meilleurs résultats sont obtenus avec des paires de langues plus éloignées mais mieux dotées.

Le choix du niveau de représentation des mots est une donnée importante. Nguyen & Chiang (2017) ont montré que l'utilisation de symboles sous-lexicaux partagés entre les langues des modèles parent et enfant, permettent d'augmenter les performances du transfert. Nous utilisons cette méthode en explorant différentes quantités de symboles (i.e. différentes tailles de vocabulaire).

Les travaux présentés dans cet article étendent ceux de Kocmi & Bojar (2018) sur plusieurs points. Tel Kocmi & Bojar (2018), nous cherchons à évaluer les performances du système enfant en fonction des données utilisées dans le système parent, selon les critères de proximité de langue et de quantité de données. Dans cette étude, nous considérons également un système parent constitué d'un encodeur universel (entraîné sur plusieurs langues). Nous nous interrogeons sur les différents choix à effectuer pour les pré-traitements des données et les paramètres du modèle de traduction et nous tenterons de déterminer la meilleure configuration.

L'objectif est de mieux comprendre la corrélation entre l'impact de la quantité de données et celui de

la proximité des langues sur les performances du système enfant. Nous verrons que nos expériences contredisent certaines conclusions des articles précédemment cités.

3 Données

Notre objectif est d'avoir les meilleurs résultats possibles pour la paire de langue estonien-anglais. Nous disposons de 2.5 millions de phrases parallèles pour cette paire de langue. Bien qu'on ne puisse pas considérer cela comme une paire de langue sous dotée, cette quantité reste faible pour apprendre un système obtenant de bons résultats.

Nous allons utiliser l'apprentissage par transfert, nécessitant des paires de langues additionnelles. Nous utilisons les données présentées dans la campagne d'évaluation de traduction automatique WMT2018 (Bojar *et al.*, 2018).

Pour évaluer l'impact de la proximité des langues dans le système parent, nous utilisons deux paires de langues différentes. L'une est une paire de langues proche ; nous avons choisi la paire finnois vers anglais car cette langue est proche de l'estonien, ce sont toutes deux des langues finno-ougriennes. Nous disposons de 5 millions de phrases parallèles en finnois-anglais. L'autre paire de langues que nous avons choisie est l'allemand vers l'anglais : l'allemand est une langue germanique plus éloignée du finnois et de l'estonien. En revanche, pour la paire allemand-anglais nous disposons de 40 millions de phrases parallèles, ce qui constitue un corpus de choix. Nous voulons découvrir si cette différence significative de quantités permettra à un système parent allemand-anglais de fournir un transfert au système enfant aussi efficace qu'avec un système parent finnois-anglais.

3.1 Pré-traitement de données

Afin de préparer nos données nous passons par plusieurs phases de pré-traitement des corpus. Nous utilisons des unités sous-mots SPM (Kudo & Richardson, 2018). Les systèmes utilisant des unités sous-mots forment l'état de l'art actuel en traduction automatique neuronale. Cela nous permet aussi un transfert plus important entre le parent et l'enfant corrélé au nombre de sous-mots en commun (Nguyen & Chiang, 2017).

Deux modèles d'unités SPM séparés sont appris. Le premier sur les langues sources mises en jeu dans les systèmes parent et enfant, et le second sur la langue cible (anglais). Les deux vocabulaires source et cible correspondant sont créés à partir des données tokenisées à l'aide des modèles précédents.

En entrée de nos systèmes, les langues changent lorsque nous passons de l'apprentissage du système parent à celui de l'enfant. Nous prenons cela en compte en entraînant des modèles de sous-mots pour la partie source avec les données utilisées pour apprendre le système parent et enfant. Le but est de ne pas avoir à modifier le vocabulaire lors de la transition parent/enfant. Afin de ne pas bruiter notre système, nous retirons les phrases de moins de 3 sous-mots et de plus de 100 sous-mots.

Et enfin, nous ne conservons dans nos vocabulaires que les unités sous-mots apparaissant au moins 5 fois dans nos corpus d'entraînement et faisons correspondre les autres à une unité inconnue : <unk>. Ce traitement est nécessaire dans notre cas, puisque SPM ne peut garantir la couverture exhaustive d'un corpus. Nous n'utilisons pas de tags comme dans Johnson *et al.* (2016) pour favoriser le transfert entre les langues proches car, du fait de leur proximité, ces dernières, partagent nécessairement des

sous-mots.

4 Architecture

Pour réaliser nos expériences nous nous somme basés sur le principe d'apprentissage par transfert trivial de Kocmi & Bojar (2018). Le principe est d'utiliser une architecture qui ne change pas entre l'apprentissage du système parent et du système enfant. Seules les données d'apprentissage sont changées pour passer de l'apprentissage du parent à l'enfant. Nous utilisons une architecture bout en bout de type encodeur/décodeur standard avec mécanisme d'attention en traduction automatique (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014). Cette architecture est composée d'un encodeur bi-directionnel et d'un décodeur à base d'unités récurrentes à portes, Bi-GRU (Cho *et al.*, 2014) de taille 800. Les plongements lexicaux (*embeddings*) sont de taille 400. Nous appliquons un *dropout* (Srivastava *et al.*, 2014) de 0.3 sur les *embeddings*, sur le contexte avant qu'il soit fourni au mécanisme d'attention et sur la sortie avant le *softmax*. Nous utilisons Adam (Kingma & Ba, 2014) pour optimiser les poids. Les poids sont initialisés d'après He *et al.* (2015). Le taux d'apprentissage est initialisé à 1.10e-4 et la taille d'un batch est de 32. Cette architecture est la seule configuration utilisée pour tous les systèmes présentés dans cet article. Ils ont été implémentés avec le toolkit nmtpytorch [1] (Caglayan *et al.*, 2017).

5 Expérimentations

Tous les résultats des systèmes présentés ici sont calculés sur les corpus de développement de la tâche de traduction de *news* de la campagne d'évaluation WMT2018.

Les résultats des systèmes de base ET-EN présentés dans la table 1 serviront de comparaison aux résultats provenant d'un apprentissage par transfert des systèmes enfants. Les écarts de résultats sont faibles et peu significatifs. À noter que le nombre d'unités SPM employé côté source et côté cible est identique.

Quantité d'unités SPM	ET-EN 2.5M	ET-EN 200k
8k	14.12	10.69
16k	14.17	10.70
32k	13.60	10.10

TABLE 1 – Résultats en %BLEU pour la paire de langue ET-EN sans apprentissage par transfert avec des vocabulaires comprenant seulement des sous-mots provenant des corpus d'entraînement ET coté source et EN coté cible.

Pour l'apprentissage par transfert, nous conservons le vocabulaire entre les systèmes parents et enfants. De ce fait, nous conservons aussi les modèles SPM. La quantité d'unités sous-mots que nous choisissons pour ces modèles a un impact sur la qualité des systèmes appris. Nous pouvons voir dans la table 2, que les modèles fondés sur les unités SPM comprenant de l'allemand obtiennent de moins

1. https ://github.com/lium-lst/nmtpytorch

bons résultats que ceux comprenant du finnois pour l'apprentissage d'un système estonien-anglais. Le finnois et l'estonien étant des langues proches, il est vraisemblable qu'ils partagent plus de mots qu'avec l'allemand, ce qui explique les résultats obtenus. De ce fait, ils cohabitent mieux dans le vocabulaire. Cela est confirmé par les résultats du modèle SPM estonien-allemand qui augmentent lorsque le nombre de sous-mots augmente. Alors que pour le SPM estonien-finnois les résultats baissent lorsqu'on utilise 32k unités comparées aux 16k unités précédentes. Il semble donc qu'un plus grand nombre de sous-mots soit plus propice pour le système allemand-estonien alors que 16k unités suffisent pour le système finnois-estonien.

Quantité d'unités SPM	DE+ET 2.5M	DE+ET 200k	FI+ET 2.5M	FI+ET 200k
8k	10.64	-	14.47	-
16k	11.55	9.27	15.08	10.66
32k	12.52	-	13.87	-

TABLE 2 – Résultats en %BLEU pour la paire de langue ET-EN sans apprentissage par transfert avec des vocabulaires comprenant des unités SPM relatives aux paires des systèmes parents.

Tout d'abord, les résultats des systèmes de base dans la table 3 nous donnent une idée des performances obtenues par les systèmes parents dans leurs paires de langues respectives. On observe que les performances du système parent DE-EN utilisant seulement 5M de données sélectionnées aléatoirement sont très inférieures à celles du système utilisant toutes les données disponibles. On peut donc s'attendre à une perte de performance lorsque le système parent est entraîné avec une plus faible quantité de données.

Afin non seulement d'avoir des résultats comparables pour nos systèmes enfants, mais aussi pour respecter le principe de l'apprentissage par transfert trivial, nous avons dû nous limiter à une seule configuration d'architecture. Cette dernière est celle décrite précédemment.

Pour définir de la taille de l'architecture nous avons fait un compromis afin d'avoir une taille assez grande pour apprendre correctement le système parent mais raisonnable pour ne pas sur-apprendre lors de l'apprentissage de l'enfant.

Pour la suite des expériences, nous avons choisi d'utiliser 16k sous-mots car c'est avec cette quantité que nous obtenons les meilleures performances en ET-EN.

Paire de langue	40M	5M
FI-EN	-	18.03
DE-EN	20.41	11.11

TABLE 3 – Résultats des modèles parents de base en %BLEU.

Dans la table 4, nous présentons les résultats des systèmes ET-EN enfants qui ont été appris sur une base des différents systèmes parents présents dans la table 3.

Nos systèmes montrent une amélioration face au 14.17 %BLEU du système de base ET-EN (voir table 1). Les résultats de ces systèmes sont proches mais nous voyons que, de base, les résultats avec le SPM DE+ET sont moins bons. Au final, le meilleur résultat est obtenu avec le transfert du système

FI-EN.

Nous avons utilisé 5M de phrases parallèles extraites aléatoirement des 40M dont nous disposions pour créer un corpus réduit en allemand. Ce corpus nous permet d'entraîner un système parent et de comparer les performances du transfert avec celui du système parent finnois. Les résultats montrent qu'à quantité de données équivalentes, les résultats diffèrent grandement. Nous expliquons cet écart par la proximité des langues utilisées pour entraîner le système parent. Le finnois, plus proche de l'estonien, offre un meilleur transfert que l'allemand qui est plus éloigné. Kocmi & Bojar (2018) montre que la qualité du système parent est importante pour assurer un bon transfert à un enfant. Les performances plus faibles du parent DE-EN utilisant 5M de données sont une explication possible aux faibles résultats du système enfant appris ensuite.

Nous avons aussi essayé de combiner la proximité du finnois à l'estonien et de profiter de la grande quantité de données provenant de la paire allemand-anglais. Pour cela, nous avons réalisé un système avec encodeur et décodeur universels (Ha *et al.*, 2016) avec le corpus finnois et allemand en source de notre système. Le système universel nous permet de modéliser une ou plusieurs langues dans le système sans avoir à faire évoluer l'architecture. Ainsi, nous obtenons des systèmes enfants estonien-anglais vraiment comparables tout en ayant un système multilingue comme parent. De plus, Johnson *et al.* (2016) a montré que l'apprentissage en parallèle de plusieurs paires de langues avec une architecture universelle a un impact positif sur les résultats de traduction, notamment pour les paires de langues dotées d'une quantité de données réduite. Nous voulons vérifier si c'est aussi le cas pour l'apprentissage par transfert. L'idée avec ce système multilingue est d'assembler les deux caractéristiques les plus importantes pour l'apprentissage par transfert, à savoir la proximité des langues impliquées et la quantité de données disponible. Nous utilisons donc un modèle SPM différent des précédents car il comporte cette fois-ci de l'allemand et du finnois provenant du système parent, en plus de l'estonien du système enfant pour le coté source.

L'hypothèse est qu'en combinant ces deux facteurs nous devrions obtenir un parent qui procurera un meilleur transfert à nos systèmes enfants. Les résultats nous montrent que cela n'est pas aussi évident (cf. table 4) ; les performances sont moins bonnes qu'avec l'allemand-anglais ou que le finnois-anglais comme seul parent. Une explication est que le déséquilibre des quantités de données entre les deux langues source du parent est un obstacle à l'apprentissage d'un parent de bonne qualité. Nous envisageons comme travail futur différentes répartitions de données afin de laisser plus de place au finnois dans le système parent.

Paire de langue	45M (40M+5M)	40M	5M
FI-EN	-	-	16.55
DE-EN	-	16.10	10.92
FI+DE-EN	15.71	-	-

TABLE 4 – Résultats en %BLEU des modèles enfants ET-EN avec les différents systèmes parents.

Nous constatons une amélioration grâce au transfert pour le système enfant ET-EN (cf. table 4. Toutefois, nous avons aussi voulu appliquer ce transfert dans le cadre de l'apprentissage d'un enfant où peu de ressources sont disponibles. Nous avons donc simulé ce manque de données en ne prenant que 200k phrases du corpus ET-EN pour apprendre de nouveaux enfants avec toujours les mêmes systèmes parents. Les résultats de la table 5 montrent que lorsque peu de données sont disponibles

pour le système enfant, la proximité des langues est plus importante. Le comportement à quantités de données équivalentes ne change pas lors de la comparaison du parent DE-EN à celui FI-EN. Le système parent DE-EN offre un moins bon transfert que le parent finnois. Le système parent FI-EN surpasse clairement les autres dans cette configuration. Il n'y a pas de changement pour notre parent multilingue qui donne toujours un transfert moins performant que les autres.

Paire de langue	45M (40M+5M)	40M	5M
FI-EN	-	-	13.03
DE-EN	-	11.12	7.10
FI+DE-EN	11.05	-	-

TABLE 5 – Résultats en %BLEU des modèles enfants ET-EN avec 200k phrases avec les différents systèmes parents.

6 Conclusion

Nous avons présenté une analyse de l'apprentissage par transfert pour la traduction automatique neuronale. Cette analyse contient des expériences se concentrant sur deux aspects importants du transfert, à savoir la quantité de données et la proximité des langues. L'objectif est de déterminer lequel de ces deux facteurs est le plus pertinent pour l'apprentissage par transfert. Nous avons montré que les quantités de données et la proximité des langues ont un impact dès la réalisation des unités sous-mots et des vocabulaires. Ces paramètres sont donc à prendre en compte pour le choix des systèmes parents.

Nos résultats vont dans le sens de ceux obtenus par Zoph *et al.* (2016) et Dabre *et al.* (2017) ; la proximité des langues utilisées pour l'apprentissage par transfert est un critère plus important que la quantité de données. À quantités de données équivalentes, les systèmes parents utilisant des paires de langues proches obtiennent de meilleurs résultats. Il ne faut pas, en revanche, négliger la qualité des systèmes parents en question et prendre cela en compte dans les résultats des systèmes enfants. Cette hypothèse est vérifiée lorsque nous avons une quantité "raisonnable" de phrases parallèles pour apprendre un système de traduction et lorsque nous avons peu de données. Notre approche de système universel combinant une grande quantité de données et des données plus proches n'a pas surpassé les approches classiques. Dans le futur, nous aimerions pousser cette approche en essayant différentes combinaisons de quantités de données pour mieux comprendre l'importance de la répartition des langues dans cette approche universelle.

Remerciements

Ce travail a été effectué dans le cadre du projet CHIST-ERA M2CR, financé par l'Agence Nationale de la Recherche (ANR) sous le contrat numéro ANR-15-CHR2-0006-01.

Références

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, **abs/1409.0473**.

BOJAR O., FEDERMANN C., FISHEL M., GRAHAM Y., HADDOW B., HUCK M., KOEHN P. & MONZ C. (2018). Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2 : Shared Task Papers*, p. 272–307, Belgium, Brussels : Association for Computational Linguistics.

CAGLAYAN O., GARCÍA-MARTÍNEZ M., BARDET A., ARANSA W., BOUGARES F. & BARRAULT L. (2017). Nmtpy : A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, **109**, 15–28.

CHO K., VAN MERRIENBOER B., GÜLÇEHRE Ç., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, **abs/1406.1078**.

DABRE R., NAKAGAWA T. & KAZAWA H. (2017). An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, p. 282–286 : The National University (Phillippines).

GU J., WANG Y., CHEN Y., LI V. O. K. & CHO K. (2018). Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 3622–3631, Brussels, Belgium : Association for Computational Linguistics.

HA T., NIEHUES J. & WAIBEL A. H. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, **abs/1611.04798**.

HE K., ZHANG X., REN S. & SUN J. (2015). Delving deep into rectifiers : Surpassing human-level performance on imagenet classification. *CoRR*, **abs/1502.01852**.

JOHNSON M., SCHUSTER M., LE Q. V., KRIKUN M., WU Y., CHEN Z., THORAT N., VIÉGAS F. B., WATTENBERG M., CORRADO G., HUGHES M. & DEAN J. (2016). Google's multilingual neural machine translation system : Enabling zero-shot translation. *CoRR*, **abs/1611.04558**.

KINGMA D. P. & BA J. (2014). Adam : A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

KOCMI T. & BOJAR O. (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation : Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, p. 244–252.

KOEHN P. & KNOWLES R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, p. 28–39, Vancouver : Association for Computational Linguistics.

KUDO T. & RICHARDSON J. (2018). Sentencepiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, **abs/1808.06226**.

LAKEW S. M., EROFEEVA A., NEGRI M., FEDERICO M. & TURCHI M. (2018). Transfer learning in multilingual neural machine translation with dynamic vocabulary.

NGUYEN T. Q. & CHIANG D. (2017). Transfer learning across low-resource, related languages for neural machine translation. *CoRR*, **abs/1708.09803**.

SRIVASTAVA N., HINTON G., KRIZHEVSKY A., SUTSKEVER I. & SALAKHUTDINOV R. (2014). Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**, 1929–1958.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, **abs/1409.3215**.

ZOPH B., YURET D., MAY J. & KNIGHT K. (2016). Transfer learning for low-resource neural machine translation. *CoRR*, **abs/1604.02201**.