Vers la génération automatique de gloses pour la documentation automatique des langues

Shu Okabe, François Yvon

Vers la génération automatique de gloses pour la documentation automatique des langues

Shu Okabe[1] François Yvon[1]
(1) Université Paris-Saclay, CNRS, LISN, Bât. 508, Rue du Belvédère, F-91405 Orsay, France
shu.okabe@limsi.fr, francois.yvon@limsi.fr

RÉSUMÉ

Une étape du processus de la documentation d'une langue consiste à annoter des énoncés recueillis sur le terrain – après enregistrement et transcription phonétique – au niveau des morphèmes. Concrètement, pour chaque unité minimale segmentée dans la séquence d'entrée, il s'agit d'attacher soit une (plus rarement) plusieurs étiquettes morphosyntaxiques, soit une étiquette de concept, le plus souvent représenté par le mot anglais correspondant. Dans la perspective d'automatiser cette phase d'annotation, nous présentons les résultats d'une étude préliminaire où nous la considérons comme une tâche d'étiquetage de séquences, dont nous chercherons à estimer la difficulté, en la comparant à une tâche d'étiquetage morphosyntaxique standard. La question principale qui nous anime étant d'évaluer la faisabilité de cette annotation lorsque les données d'apprentissages sont très limitées.

ABSTRACT

Towards Automatic Gloss Generation for Computational Language Documentation.

One step of the language documentation process consists in annotating utterances collected on the field—once transcribed—at the morpheme level. For each minimal unit segmented in the input stream, the annotation process associates either one (or, in rare cases, several) morpho-syntactic tag(s) or one conceptual label represented by the corresponding English lemma. With the goal of automating this annotation task, we report the results of a preliminary study where this task is viewed as a sequence labelling task. Comparing the obtained results with a standard PoS task on the same data allows us to assess the difficulty of the process, especially in the context where the training resources are limited.

MOTS-CLÉS : génération de gloses interlinéaires, documentation automatique des langues.

KEYWORDS: interlinear gloss generation, computational language documentation.

1 Introduction

La documentation automatique des langues s'intéresse aux méthodes et outils destinés à assister les linguistes de terrain dans leurs tâches de collecte et d'annotation de données linguistiques, comme illustré Figure 1. Une fois transcrit phonétiquement, un énoncé (S) est segmenté en mots (séparés par les espaces) et morphèmes (séparés par des tirets). La glose interlinéaire (G) associe à chaque morphème une étiquette correspondant soit à son sens *lexical* exprimé par un concept de la langue cible (c'est le cas de *man* sur la figure), soit à sa fonction *grammaticale* (DEF dans l'exemple de la figure). Dans cet exemple, la dernière ligne (T) correspond à la traduction de la phrase dans une langue cible, ici l'anglais.

S	Phrase segmentée	bečedaw–ni	žek'u	razi	oq–n
G	Glose	wealthy–DEF	man	agree	become–PST.UNW
T	Traduction (EN)	*The wealthy man agreed.*			

FIGURE 1 – Strates d'annotation, extraites d'un corpus en langue tsez (Abdulaev & Abdulaev, 2010)

Notre objectif ultime est de produire automatiquement les gloses à partir de la phrase segmentée dans la langue source et de sa traduction en langue cible, deux ressources systématiquement disponibles. Les gloses constituent une strate d'annotation essentielle dans l'optique d'élaborer une grammaire ou un dictionnaire. Néanmoins, cette annotation est délicate et coûteuse à réaliser ; son automatisation, même partielle, permettrait d'accélérer cette étape. Le travail exploratoire présenté ici vise à évaluer la difficulté de cette tâche, par comparaison à la tâche d'étiquetage en partie du discours (PoS), dans un cadre de documentation des langues. Ce contexte implique des corpus de taille réduite, et des jeux d'étiquettes très grands, potentiellement « ouverts » (puisque les étiquettes lexicales, contrairement aux étiquettes grammaticales, peuvent prendre la forme d'un lemme quelconque du lexique cible), qui sont deux facteurs de complexité. En revanche, la réalisation d'une annotation au niveau des morphèmes contribue à limiter la diversité des unités en source, ce qui pourrait, au contraire, être un facteur facilitant.

2 Méthodologie

Modèles Pour développer un premier système, nous avons choisi de décomposer la tâche de génération de gloses en deux étapes : (1) génération d'une série d'étiquettes de gloses grammaticales (Premières Étiquettes, PE), n'utilisant que la phrase source (S) ; (2) génération des gloses restantes en s'appuyant sur la traduction (T).

L'étape (1) est effectuée avec un champ aléatoire conditionnel (*Conditional Random Field*, CRF) (Lafferty *et al.*, 2001), en utilisant Wapiti (Lavergne *et al.*, 2010). Afin de disposer d'un ensemble fini d'étiquettes à cette étape, toutes les gloses lexicales sont unifiées sous la même étiquette, « stem », suivant la méthodologie de (Moeller & Hulden, 2018; Barriga Martínez *et al.*, 2021).

L'étape (2) consistera alors à prédire les gloses lexicales pour spécialiser l'étiquette « stem » à partir de la traduction, par exemple en utilisant des alignements automatiques. Elle n'est pas traitée dans cette étude.

Ressources linguistiques Notre première langue d'étude est le tsez, une langue nakho-daghestanienne, en utilisant les annotations extraites de (Abdulaev & Abdulaev, 2010). Ce corpus a déjà été utilisé dans (Zhao *et al.*, 2020), qui aborde la même tâche avec une approche entièrement neuronale. Il comporte 2 000 phrases glosées et traduites en anglais, comme dans l'exemple 1.

Dans un second temps, nous avons également étudié le zaar [1], une langue tchadique parlée au Nigéria, à travers un corpus doublement annoté pour chaque morphème en glose et partie du discours (Caron, 2015). Nous dénombrons dans cette ressource 190 étiquettes grammaticales (sans stem) et 29 PoS.

1. Disponible sur https://github.com/surfacesyntacticud/SUD_Zaar-Autogramm/tree/master/CORPUS_PREANNOTE.

La figure 2 présente un exemple de phrase zaar avec ses deux types d'annotations au niveau des morphèmes, la strate P représentant ici les parties du discours.

S	Phrase segmentée	tò	kəndá	tɔ́	ndâː	lɔ́ːr–kɔ́nì	səmbɔ́r–sə	dì
G	Glose	well	then	3PL.AOR	start	bring–NMLZ	stranger–PL	CTP
P	Partie du discours	PART	ADV	AUX	VERB	VERB–SCONJ	NOUN–DET	PART
T	Traduction (EN)	*Well then they started bringing guests.*						

FIGURE 2 – Extrait du corpus en langue zaar (Caron, 2015)

Le tableau 1 présente le nombre de phrases (N_{sent}), le nombre d'occurrences (N_{token}) et de types (N_{type}) de *morphèmes*, le nombre de gloses grammaticales (N_{gram} ; sans « stem ») et d'étiquettes PoS (N_{PoS}) pour ces deux corpus.

	N_{sent}	N_{token}	N_{type}	N_{gram}	N_{PoS}
Tsez	2000	40229	1603	158	/
Zaar	1707	16957	1690	190	29

TABLE 1 – Statistiques générales des corpus tsez et zaar

3 Résultats préliminaires

Nous présentons une expérience pour chaque langue. Pour le tsez, nous évaluons la tâche (1) directement, tandis que les étiquettes en partie du discours permettent de comparer les tâches d'étiquetage en gloses et en PoS pour le zaar.

Dans les deux expériences, nous séparons le corpus en trois parties et faisons varier le nombre d'énoncés pour l'apprentissage, en maintenant 200 phrases pour le développement et 200 pour le test.

Enfin, comme nous avons accès à une segmentation en mots et en morphèmes pour les deux langues, le texte segmenté en entrée du CRF intègre les tirets dans la représentation graphique. Ceci permet indirectement d'exprimer une information de position sur les morphèmes en distinguant notamment les unités qui sont positionnées au début des mots de celles qui sont à l'intérieur des mots (elles débutent dans ce cas par le caractère « - »). Ce choix a été motivé par une expérience en amont pour laquelle cette information n'était pas présente et qui a conduit à des systématiquement moindres que ceux présentés ci-dessous.

Conditions expérimentales Nous avons utilisé le paramétrage par défaut de Wapiti [2]. L'algorithme d'optimisation utilisé (par défaut) est OWL-QN (*Orthant-Wise Limited-memory Quasi-Newton*, Andrew & Gao 2007), qui semble être la meilleure approche dans des situations avec peu de données (Lavergne *et al.*, 2010). Nous présentons principalement les résultats moyennés de deux lancers avec leur écart relatif, en considérant deux jeux de données ré-échantillonnés.

2. Détaillé sur https://wapiti.limsi.fr/manual.html.

3.1 Génération de gloses en tsez

Notre première expérience se concentre sur l'étape (1), l'étiquetage des gloses par un CRF. Dans le tableau 2, nous avons suivi l'évolution du score de correction [3] (*accuracy*) en fonction de la taille des données d'entraînement. À titre indicatif, (Zhao *et al.*, 2020) obtient, avec environ 1 600 phrases d'entraînement en tsez [4], une *accuracy* de 84 % avec le modèle statistique de (McMillan-Major, 2020) et 87 % avec leur modèle neuronal, sur la tâche de génération de gloses *entière* (donc 1 + 2 selon notre décomposition).

Deux patrons ont été comparés pour le CRF : l'un ne prend en compte que les unigrammes d'étiquettes, l'autre intégrant aussi des bigrammes, tous deux sur une fenêtre de cinq mots. Nous reportons aussi les résultats obtenus avec deux autres systèmes : `stem`, qui prédit toujours l'étiquette la plus fréquente, « stem », et `maj`, qui utilise les données d'entraînement pour obtenir l'étiquette majoritaire d'un morphème et prédit toujours cette étiquette pour toutes les occurrences de la base de test (« stem » est l'étiquette par défaut qui est utilisée pour tous les mots inconnus).

Taille entraînement	200	500	800	1000	1300	1600
`stem`	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)	47,0 (0,6)
`maj`	83,6 (0,0)	84,0 (0,2)	84,0 (0,5)	84,1 (0,4)	84,1 (0,4)	84,2 (0,4)
Unigramme	**84,5** (1,7)	89,3 (1,1)	90,7 (0,0)	91,3 (0,3)	91,5 (0,8)	92,5 (0,1)
Bigramme	84,3 (1,7)	**89,7** (1,2)	**90,8** (0,3)	**91,6** (0,3)	**92,1** (0,4)	**92,8** (-)

TABLE 2 – Évolution de l'*accuracy* en fonction du nombre de phrases dans les données d'entraînement dans le corpus tsez (moyenne de deux lancers (écart type)).

Le système `stem` témoigne de la prépondérance de l'étiquette « stem » dans les données, représentant environ la moitié des étiquettes de référence. Le système `maj` quant à lui stagne autour de 84, ce qui le rend moins intéressant avec des données d'entraînement plus important.

Hormis des résultats similaires pour les deux patrons, nous pouvons voir qu'avec seulement 200 phrases d'entraînement, le modèle parvient à atteindre une correction élevée, avec en particulier un F-score de plus de 91 % pour l'étiquette « stem » (qui est l'étiquette majoritaire).

Taille entraînement	200	500	800	1000	1300	1600
`maj`	97,3	97,1	97,1	97,1	97,1	97,1
Unigramme	97,4	97,4	97,3	97,3	97,3	97,3
Bigramme	97,3	97,4	97,3	97,3	97,3	97,4

TABLE 3 – Évolution du taux de d'étiquettes « stem » correctement prédites en fonction du nombre de phrases tsez d'entraînement (résultat d'un lancer). Le système `stem` obtient par définition 100.

Le tableau 3 se concentre sur la prédiction de l'étiquette « stem » parmi les étiquettes de référence. Tous les systèmes obtiennent autour de 97 %, quel que soit le nombre de phrases en entraînement.

3. Proportion d'étiquettes correctement prédites.
4. Le nombre total de phrases dans leur corpus tsez est de 1 782 phrases.

Distinguer les étiquettes lexicales de celles grammaticales semblerait donc relativement accessible, avec peu de données.

Pour vérifier cette intuition, dans une expérience complémentaire, nous avons unifié toutes les étiquettes grammaticales sous le label « gram » et toutes les autres sous le label « stem ». Pour cette expérience, toujours pour le tsez, nous obtenons plus de 95 % d'*accuracy*.

3.2 Difficulté de la tâche par rapport à l'étiquetage en PoS

Avec la même méthodologie que pour la première expérience, nous évaluons aussi l'étiquetage en glose et en partie du discours pour le zaar. Les résultats sont présentés dans le tableau 4.

Taille	207 PoS / Gloses	507 PoS / Gloses	807 PoS / Gloses	1007 PoS / Gloses	1307 PoS / Gloses
`stem`	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)	- / 44,7 (1,8)
`maj`	**71,6** (12,7) / **83,7** (9,8)	**80,5** (5,7) / **87,9** (5,1)	**83,9** (4,0) / **88,8** (4,5)	**85,1** (3,6) / **89,0** (4,2)	86,2 (3,7) / **91,1** (1,2)
Unigramme	61,8 (8,1) / 67,8 (0,3)	77,7 (4,2) / 79,3 (1,9)	82,1 (2,4) / 82,1 (2,3)	83,5 (3,7) / 83,8 (3,2)	85,7 (2,2) / 86,4 (1,2)
Bigramme	61,8 (8,6) / 67,9 (0,1)	77,6 (4,7) / 79,0 (2,6)	82,2 (2,4) / 82,3 (2,3)	83,1 (4,0) / 83,6 (3,4)	**86,2** (2,3) / 85,3 (-)

TABLE 4 – Évolution de l'*accuracy* en fonction du nombre de phrases dans les données d'entraînement dans le corpus zaar (moyenne de deux lancers (écart type)).

En ce qui concerne les deux patrons de CRF, nous observons des tendances similaires pour la génération de gloses par rapport à l'expérience précédente. Nous notons toutefois des *accuracy* plus faibles par rapport au tsez, sans doute explicables par le nombre réduit (d'occurrences) de morphèmes dans le corpus zaar. Cette différence peut aussi expliquer les performances du système `maj`, bien meilleur lorsqu'il y a peu de données (environ 15 points de différence avec 207 phrases d'entraînement).

De plus, nous pouvons voir que l'étiquetage en gloses obtient de meilleurs scores de correction par rapport à celui en PoS, bien que très proches. La différence est plus notable lorsque la taille de données est moindre. Malgré les facteurs de complexité énoncés, la tâche de génération de gloses semble donc de difficulté comparable à l'étiquetage en parties du discours.

4 Conclusions et perspectives

Les premiers résultats de génération de gloses sur des langues peu dotées semblent encourageants, avec une *accuracy* qui dépasse 80 % pour seulement 400 phrases à l'entraînement en tsez. Diverses perspectives sont envisagées : (a) étudier la difficulté de l'étape (2), qui semble comparativement plus simple ; (b) étendre le modèle par des ressources externes (dictionnaires, phrases non étiquetées) ; (c) construire des modèles de génération intégrés capables de réaliser simultanément les deux étapes.

Remerciements

Ce travail est effectué dans le cadre du projet franco-allemand « La documentation automatique des langues à l'horizon 2025 » (*Computational Language Documentation by 2025*, CLD 2025, ANR-19-CE38-0015-04). Les auteurs remercient Bernard Caron pour la mise à disposition du corpus zaar et Antonios Anastasopoulos pour les données tsez.

Références

ABDULAEV A. K. & ABDULAEV I. K. (2010). *Cezjas fol'klor : (gíurus mecrek°iorno butirno) = Dido (Tsez) folklore = Didojskij (cezskij) fol'klor*. Leipzig : Lotos.

ANDREW G. & GAO J. (2007). Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, p. 33–40, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1273496.1273501.

BARRIGA MARTÍNEZ D., MIJANGOS V. & GUTIERREZ-VASQUES X. (2021). Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, p. 34–43, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.americasnlp-1.5.

CARON B. (2015). Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds) (2012). 'The CorpAfroAs Corpus'. ANR CorpAfroAs : a Corpus for Afro-Asiatic languages. Document électronique. Esquisse grammaticale du zaar (langue tchadique du Nigéria), HAL : halshs-00647526.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

MCMILLAN-MAJOR A. (2020). Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, p. 355–366, New York, New York : Association for Computational Linguistics.

MOELLER S. & HULDEN M. (2018). Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, p. 84–93, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ZHAO X., OZAKI S., ANASTASOPOULOS A., NEUBIG G. & LEVIN L. (2020). Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5397–5408, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.471.