LISN @ défi partagé SIGMORPHON 2023 sur la génération automatique de gloses

Journée glose2023

Shu Okabe François Yvon

28 juin 2023

Laboratoire Interdisciplinaire des Sciences du Numérique – LISN, CNRS, Université Paris-Saclay

Introduction

T	Phrase non segmentée	Nesis	ɬˢono	uži	zown.
M	Phrase segmentée	nesi–s	ɬˢono	uži	zow–n
G	Phrase glosée	he.OBL–GEN1	three	son	be.NPRS–PST.UNW
L	Traduction (EN)	*He had three sons.*			

Figure 1 – Strates d'annotation d'une phrase en tsez (Abdulaev and Abdulaev, 2010)

T	Phrase non segmentée	Nesis	ɬˢono	uži	zown.
M	Phrase segmentée	nesi–s	ɬˢono	uži	zow–n
G	Phrase glosée	he.OBL–GEN1	three	son	be.NPRS–PST.UNW
L	Traduction (EN)	*He had three sons.*			

Figure 1 – Strates d'annotation d'une phrase en tsez (Abdulaev and Abdulaev, 2010)

Deux axes d'adaptabilité :

1. Taille des données d'entraînement ;
2. Langue cible de traduction.

T	Phrase non segmentée	Nesis	ɬˤono	uži	zown.
M	Phrase segmentée	nesi–s	ɬˤono	uži	zow–n
G	Phrase glosée	he.OBL–GEN1	three	son	be.NPRS–PST.UNW
L	Traduction (EN)	*He had three sons.*			

Figure 1 – Strates d'annotation d'une phrase en tsez (Abdulaev and Abdulaev, 2010)

Deux axes d'adaptabilité :

1. Taille des données d'entraînement ;
2. Langue cible de traduction.

⇒ Participation à l'*open track* du défi partagé SIGMORPHON 2023

Définition de notre approche

Étiquetage de séquence : avec les gloses grammaticales [2]

Ensemble des étiquettes :

$\mathcal{Y} = \{\text{LEX}\} \bigcup \{\text{ensemble des gloses grammaticales}\}$ [1].

M	Phrase segmentée	nesi–s	ɬˤono	uži	zow–n
G”	Phrase glosée	LEX–GEN1	LEX	LEX	LEX–PST.UNW

1. Plus de 100 étiquettes.
2. Méthodologie de (Moeller and Hulden, 2018; Barriga Martínez et al., 2021).

Étiquetage de séquence : avec les gloses grammaticales [2]

Ensemble des étiquettes :

$\mathcal{Y} = \{\text{LEX}\} \bigcup \{\text{ensemble des gloses grammaticales}\}$ [1].

$\rightarrow$ Étiquetage de séquence avec un champ markovien conditionnel (CRF)

M	Phrase segmentée	nesi–s	ɬˤono	uži	zow–n
G"	Phrase glosée	LEX–GEN1	LEX	LEX	LEX–PST.UNW

1. Plus de 100 étiquettes.
2. Méthodologie de (Moeller and Hulden, 2018; Barriga Martínez et al., 2021).

Notre tâche : avec les gloses lexicales

Entrée **M**	Phrase segmentée	nesi–s	ɬ$^{\text{ʕ}}$ono	uži	zow–n
Sortie **G**	Phrase glosée	he.OBL–GEN1	three	son	be.NPRS–PST.UNW

⇒ L'ensemble des étiquettes n'est plus fixé :
les gloses lexicales en nombre quasi illimité.

Notre tâche : avec les gloses lexicales

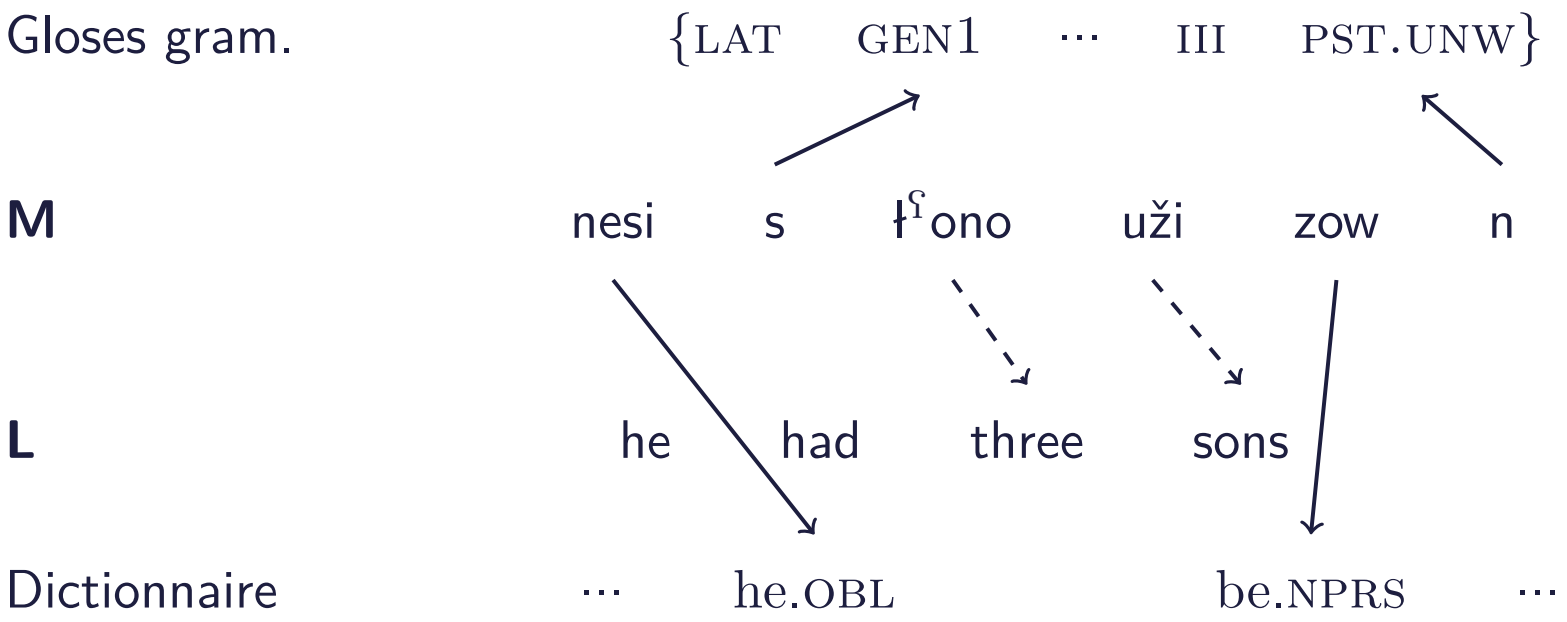
Entrée **M**	Phrase segmentée	nesi–s	ɬ$^{\text{ʕ}}$ono	uži	zow–n
Entrée **L**	Traduction (EN)	*He had three sons.*			
Sortie **G**	Phrase glosée	he.OBL–GEN1	three	son	be.NPRS–PST.UNW

⇒ L'ensemble des étiquettes n'est plus fixé :
les gloses lexicales en nombre quasi illimité.

Hypothèse :

- supposer que les gloses lexicales peuvent être déduites de la traduction

Illustration de notre approche



Glose he.OBL–GEN1 three son be.NPRS–PST.UNW

Notre approche face aux défis posés par les gloses lexicales

(a) l'emploi d'alignements déterministes obtenus avec SimAlign (Jalili Sabet et al., 2020) pour créer les étiquetages de **référence** et superviser l'apprentissage ;

(b) la spécification *locale* de $\mathcal{Y}$ (espace de recherche) en exploitant les mots de la traduction et les gloses du corpus d'apprentissage ;

(c) une implémentation basée sur Lost (Lavergne et al., 2013) qui permet de spécifier localement l'espace de recherche.

Méthodologie

Utilisation d'un modèle d'alignement : SimAlign

- SimAlign (Jalili Sabet et al., 2020), modèle d'alignement neuronal **multilingue** ;
- Méthode d'alignement `Match` : toutes les gloses lexicales auront un alignement ;
- Choix de la couche de BERT/mBERT : 0.

Utilisation d'un modèle d'alignement : SimAlign

- SimAlign (Jalili Sabet et al., 2020), modèle d'alignement neuronal **multilingue** ;
- Méthode d'alignement `Match` : toutes les gloses lexicales auront un alignement ;
- Choix de la couche de BERT/mBERT : 0.

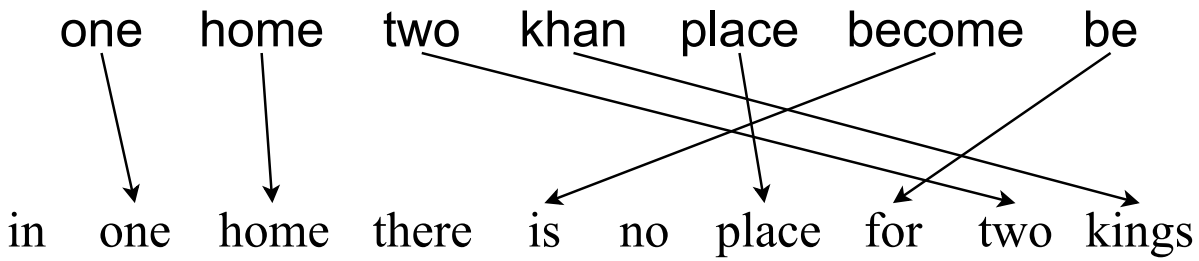


Figure 2 – Exemple d'alignement entre gloses lexicales et traduction

Constitution de l'espace de recherche

- $\mathcal{G}$: les gloses grammaticales, ensemble commun à toutes les phrases ;

Constitution de l'espace de recherche

- $\mathcal{G}$: les gloses grammaticales, ensemble commun à toutes les phrases ;
- $\mathcal{T}$: les lemmes des mots de la traduction ;

Constitution de l'espace de recherche

- $\mathcal{G}$: les gloses grammaticales, ensemble commun à toutes les phrases ;
- $\mathcal{T}$: les lemmes des mots de la traduction ;
- $\mathcal{D}$: les gloses lexicales les plus fréquemment associées aux morphèmes vus à l'entraînement (dictionnaire) ;

Constitution de l'espace de recherche

- $\mathcal{G}$: les gloses grammaticales, ensemble commun à toutes les phrases ;
- $\mathcal{T}$: les lemmes des mots de la traduction ;
- $\mathcal{D}$: les gloses lexicales les plus fréquemment associées aux morphèmes vus à l'entraînement (dictionnaire) ;
- $\mathcal{R}$: les gloses de référence, uniquement à l'entraînement

	global	local	espace de recherche
apprentissage	$\mathcal{G}$	$\mathcal{T} \cup \mathcal{D} \cup \mathcal{R}$	$\mathcal{G} \cup \mathcal{T} \cup \mathcal{D} \cup \mathcal{R}$
inférence	$\mathcal{G}$	$\mathcal{T} \cup \mathcal{D}$	$\mathcal{G} \cup \mathcal{T} \cup \mathcal{D}$

Étiquettes structurées à prédire : mieux généraliser

- GLO : la glose à prédire ;
- BIN : la nature de la glose (LEX ou GRAM) ;
- POS : l'étiquette PoS associée au mot aligné dans la traduction.

ensemble	GLO	BIN	POS
$\mathcal{G}$	GEN1	GRAM	GRAM
$\mathcal{T}$	king	LEX	NOUN
$\mathcal{D}$	khan	LEX	NOUN
$\mathcal{R}$	khan	LEX	NOUN

Caractéristiques étudiées dans les deux systèmes

Deux systèmes :

S1 : fonctions caractéristiques d'un morphème :

- sa position au sein du mot PIW,
- sa longueur LNG,
- ses trois premières et dernières lettres (PRE et SUF).

S2 : S1 + caractéristiques supplémentaires de copie et de position relative.

entrée	caractéristiques S1				S2	sorties			S2
morphème	PIW	LNG	PRE	SUF	position source	GLO	BIN	POS	position cible
nesi	0	4	nes	esi	1/4	he.OBL	LEX	PRON	1/4
s	1	1	s	s	1/4	GEN1	GRAM	GRAM	-2
ɬˤono	F	5	ɬˤo	ono	2/4	three	LEX	NUM	3/4
uži	F	3	uži	uži	2/4	son	LEX	NOUN	4/4
zow	0	3	zow	zow	3/4	be.NPRS	LEX	VERB	2/4
n	1	1	n	n	4/4	PST.UNW	GRAM	GRAM	-2

Langues étudiées

5 langues parmi 7 : tsez (ddo), gitksan (git), lezgi (lez), natugu (ntu) et uspanteko (usp).

langue	ddo	git	lez	ntu	usp
entraînement	3 558	31	701	791	9 774
développement	445	42	88	99	232
test	445	37	87	99	633
langue cible	EN	EN	EN	EN	ES

Table 1 – Nombre de phrases et langue de traduction pour chaque langue

Langues étudiées

5 langues parmi 7 : tsez (ddo), gitksan (git), lezgi (lez), natugu (ntu) et uspanteko (usp).

langue	ddo	git	lez	ntu	usp
entraînement	3 558	31	701	791	9 774
développement	445	42	88	99	232
test	445	37	87	99	633
langue cible	EN	EN	EN	EN	ES

Table 1 – Nombre de phrases et langue de traduction pour chaque langue

1. Taille de données variant de 31 à 9 774 phrases d'entraînement ;
2. Deux langues de documentation (cible).

Résultats

Résultats expérimentaux

- Métrique : exactitude (*accuracy*) au niveau des mots (haut) et des morphèmes (bas)

modèle	ddo	git	lez	ntu	usp
BASE SIG	75,7	16,4	34,5	41,1	76,6
BEST SIG	**85,8**	**31,5**	**85,4**	**89,3**	**78,5**
S1	84,9	28,4	83,4	88,8	76,3
S2	85,5	**31,5**	83,0	**89,3**	76,7
BASE SIG	85,3	25,3	51,8	49,0	82,5
BEST SIG	**92,0**	**52,4**	**87,6**	**92,8**	**84,5**
S1	91,4	50,8	87,2	92,6	82,4
S2	91,8	51,1	87,0	**92,8**	82,7

Données d'entraînement de taille variable

- Métrique : Score F1 pour les gloses grammaticales et lexicales

	S1		S2	
	gram	lex	gram	lex
200	93,3	80,5	93,6	81,3
500	95,3	88,5	95,2	88,3
entier	95,7	89,5	95,9	89,6

Table 2 – Score F1 avec des données d'entraînement de taille variable en natugu.

Conclusion

Conclusion

- Utilisation d'alignements automatiques pour superviser l'apprentissage des gloses lexicales ;
- Spécification *locale* des étiquettes grâce à un modèle basé sur les CRF, Lost ;
- Deux systèmes de fonctions caractéristiques, aux résultats compétitifs par rapport aux meilleurs systèmes du défi partagé.

Conclusion

- Utilisation d'alignements automatiques pour superviser l'apprentissage des gloses lexicales ;
- Spécification *locale* des étiquettes grâce à un modèle basé sur les CRF, Lost ;
- Deux systèmes de fonctions caractéristiques, aux résultats compétitifs par rapport aux meilleurs systèmes du défi partagé.

Perspectives :

- Amélioration des caractéristiques, extension des alignements ;
- Utilisation de méthodes neuronales.

Merci !

Références

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

Barriga Martínez, D., Mijangos, V., and 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*, pages 34–43, Online. Association for Computational Linguistics.

Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign : High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Lavergne, T., Allauzen, A., and Yvon, F. (2013). A fully discriminative training framework for statistical machine translation (un cadre d'apprentissage intégralement discriminant pour la traduction statistique) [in French]. In *Proceedings of TALN 2013 (Volume 1 : Long Papers)*, pages 450–463, Les Sables d'Olonne, France. ATALA.

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