Un corpus en arabe annoté manuellement avec des sens WordNet

Marwa Hadj Salah[1, 2] Hervé Blanchon[1] Mounir Zrigui[2] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP[*], LIG, 38000 Grenoble, France
(2) LaTICE, Tunis, 1008, Tunisie
Prénom.Nom@univ-grenoble-alpes.fr, Prénom.Nom@fsm.rnu.tn

RÉSUMÉ

OntoNotes comprend le seul corpus manuellement annoté en sens librement disponible pour l'arabe. Elle reste peu connue et utilisée certainement parce que le projet s'est achevé sans lier cet inventaire au *Princeton WordNet* qui lui aurait ouvert l'accès à son riche écosystème. Dans cet article, nous présentons une version étendue de OntoNotes Release 5.0 que nous avons créée en suivant une méthodologie de construction semi-automatique. Il s'agit d'une mise à jour de la partie arabe annotée en sens du corpus en ajoutant l'alignement vers le *Princeton WordNet* 3.0. Cette ressource qui comprend plus de 12 500 mots annotés est librement disponible pour la communauté. Nous espérons qu'elle deviendra un standard pour l'évaluation de la désambiguïsation lexicale de l'arabe.

ABSTRACT

Arabic Manually Sense Annotated Corpus with WordNet Senses

OntoNotes is the only Arabic Manually Annotated Corpus freely available for the Arabic language. It remains little known and exploited certainly because the project ended without linking this inventory to *Princeton WordNet* which would have given it access to its rich ecosystem. In this article, we present an extended version of OntoNotes Release 5.0 that we created using a semi-automatic construction methodology. This is an update of the Arabic part of the sense-annotated corpus by adding the alignment to the *Princeton WordNet* 3.0. This resource that includes more than 12,500 annotated words will be freely available for the community. We hope that it will become a standard for the evaluation of the lexical disambiguation of Arabic.

MOTS-CLÉS : Corpus annoté en sens, langue arabe, alignement de sens interlingues.

KEYWORDS: Sense annotated corpus, arabic language, interlingual sense alignment.

1 OntoNotes Release 5.0

Le projet OntoNotes (Weischedel *et al.*, 2013) est le résultat d'un travail collaboratif entre *BBN Technologies*, l'Université du Colorado, l'Université de Pennsylvanie et l'Institut des sciences de l'information de l'Université de Californie du Sud. *OntoNotes Release 5.0* est la dernière version proposée par ce projet. C'est un grand corpus annoté libre de droit, construit à 90% d'accord inter-annotateur avec des informations structurelles (syntaxe et structures prédicat-arguments) et sémantiques superficielles (sens du mot lié à une ontologie et co-référence). Le corpus contient plusieurs genres de textes en anglais et chinois et uniquement des données News pour la partie arabe.

∗. Institute of Engineering Univ. Grenoble Alpes

2 Enrichissement de la partie arabe de l'OntoNotes Release 5.0

Les parties anglaises et chinoises de *OntoNotes Release 5.0* sont annotées avec des sens issus du *Princeton WordNet*. Malheureusement, le projet n'a pas pu être mené jusqu'au bout sur la partie arabe et le lien entre les annotations *OntoNotes* et le *Princeton WordNet* sont absentes. Nous proposons ici une mise à jour de la partie arabe de *OntoNotes Release 5.0* d'une manière semi-automatique pour obtenir des mots annotés en sens avec le *Princeton WordNet* 3.0. La figure 1 présente l'architecture globale de la partie arabe de l'Ontonote.

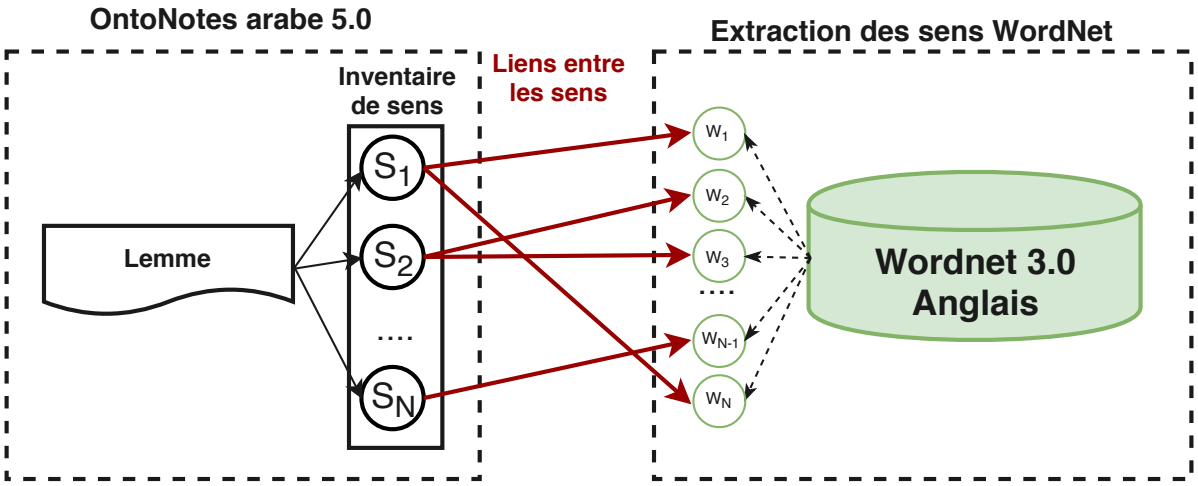

FIGURE 1 – Architecture globale de l'*OntoNotes Release 5.0* après l'ajout des correspondances pour les 261 lemmes uniques de leurs sens vers ceux du *Princeton WordNet* 3.0

Ce traitement semi-automatique d'annotations et de vérification réalisé s'est avéré coûteux en temps (quatre mois.hommes de travail). Le tableau 1 présente la description d'*OntoNotes Release 5.0* ainsi que le nombre de correspondances$_{WordNet}$ uniques ajoutées.

	#Lemmes	#Lemmes uniques	#Sens uniques	**#Correspondances$_{WordNet}$ uniques**
Verbes	3990	150	642	**4182**
Noms	8534	111	463	**1376**
Total	12524	261	1105	**5558**

TABLE 1 – Description d'*OntoNotes Release 5.0* après l'ajout des correspondances vers le *Princeton WordNet* 3.0

La version étendue de l'*OntoNote 5.0* sera disponible pour la communauté et pourra être utilisée dans plusieurs applications du traitement automatique du langage naturel pour la langue arabe, notamment dans la tâche de désambiguïsation lexicale. Avant ce travail, il n'existait aucun corpus en arabe manuellement annoté en sens pour la langue arabe qui soit librement disponible. Cette ressource facilitera la comparaison et/ou la construction de systèmes de désambiguïsation lexicale pour cette langue.

Références

F. BENARMARA, N. HATOUT, P. MULLER & S. OZDOWSKA, Eds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

G. DIAS, Ed. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benarmara *et al.*, 2007), p. 101–110.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benarmara *et al.*, 2007), p. 401–410.

WEISCHEDEL R., PALMER M., MARCUS M., HOVY E., PRADHAN S., RAMSHAW L., XUE N., TAYLOR A., KAUFMAN J., FRANCHINI M., EL-BACHOUTI M., BELVIN R. & HOUSTON A. (2013). Ontonotes release 5.0. *LDC2013T19. Web Download. Philadelphia : Linguistic Data Consortium*.