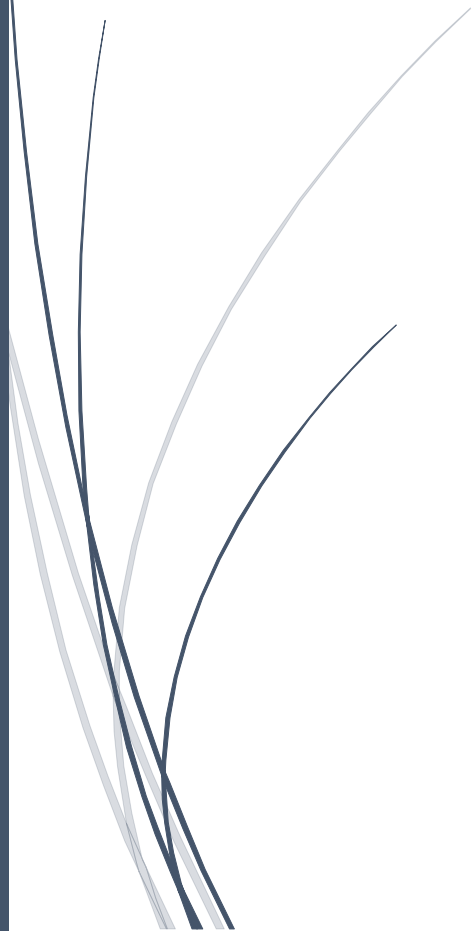
M1 Industries de la Langue

Les formats CoNLL

Synthèse technique



Jérémy Bourdillat
UNIVERSITE GRENOBLE-ALPES

Table des matières

Introduction

Le format de données CoNLL et ses dérivés est l'un des plus utilisés dans les applications de Traitement Automatique des Langues Naturelles (à tout le moins quand il s'agit de langue écrite).

Chaque variante propose un format conventionné et normé permettant à la fois l'import, l'écriture et la représentation de données linguistiques, et donc l'échange de ces données. Les applications qui utilisent ce format autorisent ainsi à l'utilisateur une certaine flexibilité dans le pipeline en favorisant l'interopérabilité entre différents modules de traitement (à condition de suivre le même format CoNLL ou de convertir).

Historique

Le format CoNLL émerge de la *Association for Computational Linguistics* (ou ACL : https://www.aclweb.org/portal/), et plus spécifiquement du *Special Interest Group on Natural Language Learning* (ou SIGNLL : https://www.signll.org/) qui vise à développer et promouvoir l'analyse automatisée du langage naturel. Cet organisme organise tous les ans, et ce depuis 1997, une conférence appelée CoNLL : *Conference on Computational Natural Language Learning*.

Depuis 1999 (soit la troisième édition de la conférence) est organisé conjointement une « tâche partagée » (*shared task*), qui est un challenge de TAL (comme on en trouve beaucoup) autour d'un des thèmes de la CoNLL, avec pour principe de fournir des données d'entraînement et de test que plusieurs équipes vont tenter de traiter du mieux possible avec un système qu'elles auront chacune mis au point et ce pour atteindre un objectif précis.

Afin d'obtenir des données de sortie comparables entre les différentes équipes et pour l'évaluation, il est utile de normer celles-ci avec un standard qui soit à la fois **simple** (à mettre en place, à lire…), **flexible** (permettant divers niveaux de détail), **léger** (en temps de calcul et en poids mémoire) et **transparent** (lisible par un humain). C'est ainsi que chaque *shared task* CoNLL a vu la nécessité de définir un format spécifique de sortie (et parfois d'entrée, pour uniformiser les données).

Quatre de ces formats se détachent par leur utilisation au-delà des frontières de ces évènements :

- **CoNLL-2003**[1], qui permet de représenter les Entités Nommées

[1] Erik F. Tjong Kim Sang et Fien De Meulder, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », in *Proceedings of CoNLL-2003*, éd. par Walter Daelemans et Miles Osborne (Edmonton, Canada, 2003), 142-47.

- **CoNLL-X**[2] (ou CoNLL-2006) qui représente les relations syntaxiques en dépendances, l'analyse lexicale et les traits morphosyntaxiques ; celui-ci a longtemps été une référence
- **CoNLL-U**[3] qui est une variation de CoNLL-X adaptée aux *Universal Dependencies* (UDs), maintenant le plus courant
- **CoNLL-U Plus**[4] qui est une version flexible et enrichie de CoNLL-U qui permet des extensions en fonction des besoins

Principales caractéristiques

Les données au format CoNLL partagent un certain nombre de caractéristiques communes. Voici les **invariants**, peu importe le format :

- Fichier texte en UTF-8[5] (pour permettre la représentation d'une majorité d'alphabets et symboles et donc de langues)
- Structure :
    - **En-tête** en début de fichier *et/ou* de phrase, avec des lignes commençant par un « # »
    - Séparateur de **colonnes** : tabulation « \t »
    - Séparateur de **lignes** : retour à la ligne « \n »
    - 1 **token** = 1 ligne
    - **Phrases** séparées par une ligne vide

Le nombre de colonnes, leurs valeurs et leur nom dépendent entièrement du format choisi.

La section suivante s'attache à décrire avec précision chacun des principaux formats CoNLL.

[2] Sabine Buchholz et Erwin Marsi, « CoNLL-X Shared Task on Multilingual Dependency Parsing », in *Proceedings of the Tenth Conference on Computational Natural Language Learning - CoNLL-X '06* (the Tenth Conference, New York City, New York: Association for Computational Linguistics, 2006), 149, https://doi.org/10.3115/1596276.1596305.

[3] « CoNLL-U Format », consulté le 19 février 2023, https://universaldependencies.org/format.html.

[4] « CoNLL-U Plus Format », consulté le 19 février 2023, https://universaldependencies.org/ext-format.html.

[5] Buchholz et Marsi, « CoNLL-X Shared Task on Multilingual Dependency Parsing », 151.