Une nouvelle méthode ensembliste pour la reconnaissance et la désambiguïsation d'entités nommées en utilisant des réseaux de neurones

Lorenzo Canale[1,2], Pasquale Lisena[1], and Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France
{canale|lisena|troncy}@eurecom.fr

[2] Politecnico di Torino, Turin, Italie

Mots-clés : Annotation sémantique, extraction d'entités nommées, désambiguïsation, apprentissage profond.

Résumé de [Canale *et al.* (2018)], publié à ISWC 2018.

Une tâche cruciale en extraction de connaissances à partir de textes se décompose souvent en deux tâches complémentaires : la reconnaissance d'entité nommée (NER) et la désambiguïsation d'entité nommée (NED). L'objectif consiste à attribuer à des parties du texte (mention) respectivement un type appartenant à une taxonomie prédéfinie et un identifiant unique, souvent représenté sous la forme d'URI, qui fait référence de manière univoque à une entité définie dans une base de connaissances donnée. La combinaison de ces deux tâches est souvent abrégée avec l'acronyme NERD.

De nombreuses approches, souvent exposées sous forme d'API Web, ont été proposées pour résoudre ces tâches au cours des dernières années. En termes de NER, chaque service fournit généralement sa propre taxonomie de types qui peuvent être reconnus. Même si tous comprendront trois types principaux (PERSON, ORGANIZATION, LOCATION), ils diffèrent largement pour les types plus fins, ce qui complique leur comparaison et leur combinaison. En termes de NED, chaque extracteur peut potentiellement lever l'ambiguïté d'entités par rapport à des bases de connaissances spécifiques (KB), mais en pratique, ils s'appuient principalement sur des bases de connaissances généralistes, comme DBpedia ou Wikidata. Pour cette raison, la comparaison et la fusion des résultats de ces extracteurs nécessitent certaines tâches de post-traitement qui dépendent généralement d'alignements entre ces bases de connaissances.

Dans ce travail, nous décrivons **Ensemble NERD**, un framework qui regroupe de nombreuses réponses d'extracteurs, les normalise et les combine afin de produire des annotations sémantiques. Cette méthode repose sur deux réseaux d'apprentissage profond, ENNTR (Ensemble Neural Network for Type Recognition) et ENND (Ensemble Neural Network for Disambiguation), qui fournissent des modèles pour effectuer d'une part un alignement entre les types et d'autre part entre les entités nommées identifiées dans une base de connaissances.

En entrée, ces réseaux reçoivent une représentation vectorielle de quatre types de caractéristiques différentes :

— Caractéristiques de forme de surface, liées au texte, à partir desquelles nous avons calculé un plongement lexical ;
— Caractéristiques de type, encodage one-hot calculé sur la taxonomie de chaque extracteur source ;
— Caractéristiques des entités, comparaison des attributs des entités extraites (étiquettes, uris, résumés, etc.)
— Caractéristiques de score, qui incluent certains scores renvoyés par les extracteurs, tels que la saillance ou la confiance.

Chaque type de caractéristique passe auparavant à travers une couche dense qui fonctionne de manière autonome par rapport aux autres, pour leur fournir une couche ensablée, qui a pour résultat la probabilité de correction d'une extraction spécifique. Cette stratégie est évaluée par rapport à des jeux de données bien connus, montrant que la production de l'ensemble surpasse les résultats obtenus par des extracteurs pris individuellement.

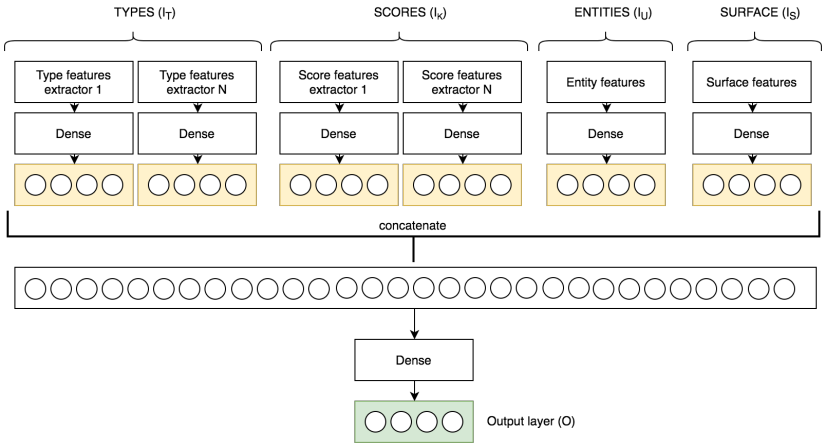

FIGURE *1 – Architecture ENNTR*

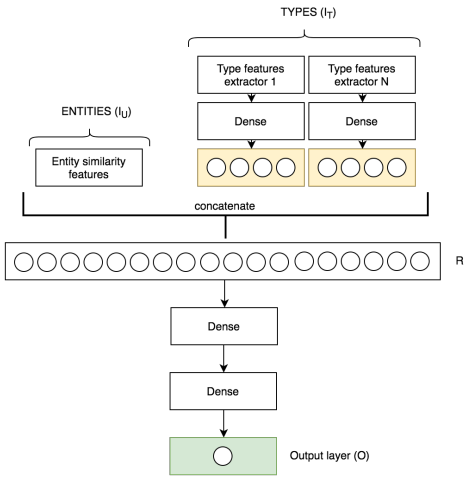

FIGURE *2 – Architecture ENND*

Les modèles produits permettent d'avoir une meilleure précision en termes de mesures micro et macro F1 par rapport aux résultats de chaque extracteur, en utilisant le framework GERBIL. De plus, le réseau ENNTR permet d'éviter l'alignement manuel entre les taxonomies de type de chaque extracteur et la taxonomie cible, en mettant en place un alignement automatique dans le premier niveau du réseau de neurones. Nous avons démontré l'importance de la sélection de caractéristiques (features) pour le succès de ces méthodes ensemblistes. En termes de NER, les formes textuelles jouent un rôle essentiel dans l'ensemble. Pour la tâche NED, si en peut bien dire que les entités aient l'impact le plus important, seule une combinaison avec leurs types permet d'améliorer réellement l'efficacité de la méthode ensembliste en ce qui concerne les prévisions produites par un seul extracteur.

Remerciements. Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet DOREMUS (ANR-14-CE24-0020), et par le programme de recherche européen H2020 dans le cadre du projet MeMAD (Subvention No. 780069).

Références

CANALE L., LISENA P. & TRONCY R. (2018). A Novel Ensemble Method for Named Entity Recognition and Disambiguation Based on Neural Network. In *17th International Semantic Web Conference (ISWC)*, Monterey, USA.