

Appendices

A Extraction d’annotations automatiques pour les données gold de la Sharedtask

A.1 Langues concernées

1. Tchèque CS: POS.
2. Français FR: POS, dépendances syntaxique.
3. Hongrois HU: POS, dépendances syntaxique.
4. Grec PL: POS.

A.2 Extraction de POS: Jackknifing

- Découper l’ensemble d’entraînement de la Sharedtask (ci-après Train) en 10 folds;
- Pour chaque fold:
 - Utiliser les folds restants comme ensemble d’entraînement d’un modèle [Marmot](#);
 - Annoter les POS du fold par ce modèle entraîné;
- Regrouper tous les folds annotés automatiquement pour créer la version automatique du Train;
- Entraîner un modèle Marmot sur Train (version gold) pour produire la version automatique de l’ensemble de test de la Sharedtask (ci-après Test).

A.3 Extraction de dépendances syntaxiques

A.3.1 CS HU et PL: Jackknifing

- Mettre le Train et le Test (version POS automatique) en format CoNLL 2009;
- Découper Train en 10 folds;
- Pour chaque fold:
 - Utiliser les folds restants comme ensemble d’entraînement d’un modèle [Mate](#);
 - Annoter les dépendances syntaxiques du fold par ce modèle entraîné;
- Regrouper tous les folds annotés automatiquement pour créer la version automatique du Train;
- Entraîner un modèle Mate sur Train (version gold) pour produire la version automatique de l’ensemble de test de la Sharedtask (ci-après Test).
- Remettre le Train et le Test en format de CoNLL standard;

A.3.2 FR

1. Regrouper les ensembles (train- val- test) de SPMRL dans un seul fichier;
2. Mettre le fichier produit en format CoNLL 2009;
3. Entraîner un modèle Mate en utilisant ce fichier comme corpus d’entraînement;
4. Utiliser ce modèle pour prédire les dépendances syntaxiques du Train;
5. Intégrer les dépendances prédites à la version de POS automatique.