Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques (ESTER)

Deuxième complément au plan d'évaluation phase 2
Procédure de soumission

Tâche Entités Nommées

Version 0.3 du 9 février 2005

1 Introduction

Ce document décrit le format de soumission pour la tâche expérimentale de détection d'entités nommées de la phase 2 de la campagne ESTER, et vient préciser le protocole d'évaluation de cette tâche, complétant le plan d'évaluation déjà diffusé le 7 janvier 2005 et comme prévu dans la section 7 de ce plan d'évaluation.

2 Description du format de soumission

2.1 Conventions de nommage

Comme lors de la soumission des résultats des tâches précédente, une expérimentation est identifiée par le nom suivant :

```
EXP-ID := <SITE>_<TACHE>_<SYS>_<DONNEES>
```

où, ici,

```
<SITE> = irisa | lia | ...
<TACHE> = en
<SYS> = primaire | contraste-xxx | contraste-yyy | ...
<DONNEES> = dev | test
```

2.2 Description des systèmes

Pour chaque expérimentation soumise, un fichier d'environ une page décrit les données, les approches algorithmiques, la configuration, les temps de calcul, etc. Il est structuré de la manière suivante :

1. `EXP-ID`

2. Description du système et des données utilisées
3. Configuration matérielle et temps de calcul
4. Références éventuelles

Pour la description du système, on veillera à détailler les points suivants :
- approche générale, nombre et type de passes
- toolkits logiciel,
- techniques et critères d'apprentissage,
- modèles,
- paramétrages.

Le fichier de description de système est nommé

```
<EXP-ID>.txt
```

2.3 Structure de l'archive

Les sorties de système à évaluer sont regroupées dans un fichier archive (tgz, tar.gz ou zip). La structure de l'archive sera la suivante :

```
<EXP-ID>.txt
<EXP-ID>/<FICHIERS-RESULTATS>
```

où le format des fichiers résultats (.ctm-ne ou .stm-ne) est décrit dans la section suivante.

3 Format des fichiers de résultats

3.1 Détection d'entités nommées dans les transcriptions de référence

Pour la tâche EN sur transcriptions de référence, les résultats seront retournés au format STM-NE (Segment Time-Marked - Named Entities). Ce format correspond à une modification du format initial STM, utilisé pour fournir les transcriptions de référence. Chaque ligne du fichier identifie un segment selon le format suivant :

```
source canal locuteur début fin <type> transcription
```

où la signification des champs est :

- `source` : nom du fichier sans extension ni chemin,
- `canal` : canal audio, toujours égal à `'1'` dans le cas présent,
- `locuteur` nom du locuteur du segment, tel que donné dans les transcriptions de référence, ou `excluded_region` dans le cas de segment ignores, ou `inter_segment_gap` dans le cas de segments vides
- `début` : date de début du segment, en secondes par rapport au début du fichier,
- `fin` : date de fin du segment, en secondes par rapport au début du fichier,

– `<type>` : type du segment,
– `transcription :` transcription orthographique.

Le champ `transcription` respecte le format suivant :

`mot1 [typeEN1 mot2 mot3 ] mot4 [typeEN2 mot5 mot6 mot7 ] mot8`

où chaque `mot[1-6]` représente un mot de la transcription, `mot2 mot3` est une première entité de type `typeEN1` et `mot5 mot6 mot7` est une seconde entité de type `typeEN2`. Les signes de ponctuation ne sont pas représentés ici dans un soucis de clarté, mais ne sont pas exclus, y compris à l'intérieur d'entités : ceux-ci seront supprimés lors de l'étape de normalisation.

Une entité ne doit pas être découpée sur plus d'un segment, et la soumission ne doit pas comporter de soumissions imbriquées (dans le cas contraire, elles seront ignorées).

Le `typeEN` doit figurer parmi la liste suivante : `pers org loc gsp fac prod time amount` .

Attention : la détection d'entités nommées ne doit pas modifier les mots de la transcription de référence.

3.2 Détection d'entités nommées dans les transcriptions automatiques

Pour la tâche EN sur transcriptions automatiques, les résultats seront retournés au format CTM-NE (Conversation Time-Mark - Named Entities). Ce format est une adaptation du format CTM décrit dans le plan d'évaluation (section A). Il respecte le format suivant :

`source canal début durée mot(--typeEN--idEN) confiance`

où les champs `source canal début durée confiance` ne diffèrent pas du format CTM initial (se référer au plan d'évaluation).

Seul le champ `mot` est modifié, dans le cas où il correspond à une entité. Dans ce cas, il est remplacé par `mot--typeEN--idEN` où `typeEN` correspond au type de l'entité (parmi `pers org loc gsp fac prod time amount`), et `idEN` correspond à un identifiant pour l'entité. L'identifiant doit être unique pour chaque instance d'entité, c'est-à-dire qu'il doit être différent pour deux références distinctes de la même entité (voir les fichiers d'exemples joints au package de scoring).

3.3 Sous-types

Une précision complémentaire du type `typeEN` (sous-type, métonymies) n'est pas exclue dans les résultats soumis mais ne sera pas prise en compte pour les évaluations de la phase 2. Elle doit alors respecter le format suivant :

`type1.soustype1/type2.soustype2`

où `soustype` vient préciser le `type`, et où le `'/'` sépare les deux types dans le cas d'une métonymie. Attention : pour la métonymie, seul le second type, `type2` , sera pris en compte.

Au niveau des références utilisées pour l'évaluation, les métonymies seront considérées comme des alternatives. Les précisions des sous-types ne seront pas pris en compte.

4 Soumission des résultats

Les soumissions se font par envoi d'un message électronique à l'adresse suivante :

`ester-soumission@etca.fr`,

avec le titre suivant

`soumission <EXP-ID>`,

et ayant en attachement l'archive contenant les sorties à évaluer.

Conformément au plan d'évaluation, les soumissions sont attendues pour le jeudi 10 février. Un accusé de réception du message sera retourné sous 48h (et dans la mesure du possible dès le vendredi 11 février). Afin de favoriser la participation à cette tâche, des soumissions tardives supplémentaires pourront être prises en compte jusqu'au 1er mars.

5 Mesure des performances

Pour la tâche de détection d'entités nommées, la principale mesure de performance sera une variante pondérée du *slot error rate* (SER) défini par Makhoul ([1]). Le principe de base est de fournir un taux d'erreur sur l'ensemble des entités de référence. La base de calcul est donc le nombre d'entités, et non pas le nombre de mots contenus dans les entités de référence.

Pour cette variante du *slot error rate*, on peut distinguer trois types d'erreurs :

- I : les insertions, qui sont des entités détectées dans l'hypothèse et qui n'ont aucun mot commun avec une entité de référence.
- D : les délétions, qui sont des entités de référence totalement manquées par le système.
- S : les entités de référence substituées, correspondant de façon incorrecte à des entités d'hypothèse.

L'ensemble S lui-même se décompose en plusieurs catégories d'erreurs : les erreurs simples (extension correcte avec type incorrecte - T, ou extension incorrecte avec type correct - E), et les erreurs complexes (type et extension incorrects). Ces dernières se décomposent encore en : une entités de référence correspondant à une seule entité d'hypothèse (TE pour type + extension incorrectes), ou une entité de référence correspondant à plusieurs entités d'hypothèse (M pour multiple).

Si on note R l'ensemble des entités de référence, taux d'erreur SER est alors défini par :

$$SER = \frac{\#I + \#D + \#S}{\#R}, \tag{1}$$

ou encore

$$SER = \frac{\#I + \#D + \#T + \#E + \#TE + \#M}{\#R}. \quad (2)$$

Dans le script de scoring utilisé, `score-EN.v0.2`, chaque type d'erreur est pondéré par un coefficient :

catégorie	insertion I	délétion D	type T	extension E	type+extension TE	multiple M
coût	1	1	0.5	0.5	0.8	1.5

D'autre part, de nombreuses questions restent en suspens pour la prise en compte de certains cas particuliers et doivent être discutées :

- **métonymies :** l'annotation des métonymies est encore sujette à discussions. Aussi, pour les références, nous proposons que l'ensemble des métonymies sera vue comme un ensemble d'alternatives pour les entités. La métonymie serait en revanche ignorée dans les fichiers d'hypothèse.
- **entités imbriquées :** comme pour les métonymies, l'annotation des entités imbriquées est sujette à discussions. Pour ces entités, l'annotation la plus significative est celle de l'entité la plus large, aussi semble-t-il préférable d'ignorer l'entité plus réduite pour cette évaluation.
- **coûts des différentes erreurs :** le coût de chaque type d'erreur doit être discuté.