Le corpus PASTEL pour le traitement automatique de cours magistraux

Salima Mdhaffar Antoine Laurent Yannick Estève
Laboratoire d'Informatique de l'Université du Mans (LIUM), Avenue Laennec, Le Mans, France
prenom.nom@univ-lemans.fr

RÉSUMÉ

Le projet PASTEL étudie l'acceptabilité et l'utilisabilité des transcriptions automatiques dans le cadre d'enseignements magistraux. Il s'agit d'outiller les apprenants pour enrichir de manière synchrone et automatique les informations auxquelles ils peuvent avoir accès durant la séance. Cet enrichissement s'appuie sur des traitements automatiques du langage naturel effectués sur les transcriptions automatiques. Nous présentons dans cet article un travail portant sur l'annotation d'enregistrements de cours magistraux enregistrés dans le cadre du projet CominOpenCourseware. Ces annotations visent à effectuer des expériences de transcription automatique, segmentation thématique, appariement automatique en temps réel avec des ressources externes... Ce corpus comprend plus de neuf heures de parole annotées. Nous présentons également des expériences préliminaires réalisées pour évaluer l'adaptation automatique de notre système de reconnaissance de la parole.

ABSTRACT

PASTEL corpus for automatic processing of lectures

PASTEL project studies the acceptability and the usability of automatic transcriptions for lectures. The aim of this project is to provide to learners a synchronously and automatically enrichment of the information that they can access during the session. This enrichment is based on automatic processing of natural language performed on automatic transcriptions. In this article, we present the annotation of lectures recorded in the context of the CominOpenCourseware project. These annotations aim to perform experiments of automatic transcription, thematic segmentation, automatic real-time matching with external resources...This corpus includes more than nine hours of annotated speech. Preliminary experiments are conducted to evaluate speech quality in our speech recognition system.

MOTS-CLÉS : Corpus, Transcription, Annotation, Segmentation Automatique, Enrichissement Automatique, Système de Reconnaissance de la Parole, Adaptation du Modèle de Langage.

KEYWORDS: Corpus, Transcription, Annotation, Automatic Segmentation, Automatic Enrichment, Speech Recognition System, Language Model Adaptation.

1 Introduction

Depuis la démocratisation des technologies de l'information et de la communication et leur interpénétration de plus en plus large et profonde avec les activités humaines, le monde de l'enseignement supérieur et de la formation pour adultes est de plus en plus interrogé par la société quant au renouvellement et à l'adaptation des pratiques pédagogiques. D'une part, les frontières entre apprentissage guidé et auto-apprentissage sont de moins en moins marquées, ce qui tend à la redéfinition du rôle de l'enseignant et de l'apprenant, et d'autre part la technologie, de plus en plus accessible, permet de diversifier et de combiner les modes d'interaction enseignant/apprenant et apprenant/apprenant.

Les technologies de reconnaissance de la parole approchent d'un niveau de maturité suffisant qui permet d'envisager de nouvelles possibilités au niveau de l'instrumentation des pratiques pédagogiques et générer de nouveaux usages. Par définition, la reconnaissance automatique de la parole vise à transcrire un énoncé produit oralement. Ces transcriptions automatiques peuvent alors être l'objet de traitements automatiques du langage naturel à l'aide de procédés généralement éprouvés sur du langage écrit.

Sur la base de transcriptions automatiques, il est possible de faciliter les échanges entre les apprenants ou entre tuteurs et apprenants, en s'appuyant sur ces transcriptions pour le développement d'outils pour la prise de notes individuelle et collaborative, ou la rédaction de compte-rendu, et pour la délimitation de zones particulièrement intéressantes du discours qui peuvent être gérées par le tuteur en direct ou a posteriori : zones traitant de notions mal comprises, demandes d'approfondissement ou d'exemple, etc.

Ainsi, le projet PASTEL a pour objectif d'explorer le potentiel de la transcription automatique en temps réel pour l'instrumentation de situations pédagogiques mixtes, où les modalités d'interaction sont présentielles ou à distance, synchrones ou asynchrones.

Différentes recherches dans la littérature ont démontré les avantages de développer des applications de traitement automatique des langues (TALN) dans le cadre pédagogique à savoir les travaux de (Ho *et al.*, 2005), (Hwang *et al.*, 2012) et (Shadiev *et al.*, 2014).

Dans cet article, nous décrivons le corpus que nous avons réalisé dans le cadre de ce projet, ainsi que des expériences préliminaires. Ce corpus composé de transcriptions manuelles du discours d'enseignement en situation de cours magistraux. Il s'accompagne d'une segmentation manuelle. Les données et les annotations seront distribuées sous licence libre à la communauté.

Ce corpus va aider au développement et à l'expérimentation d'une application dans un cadre pédagogique, va permettre l'évaluation des systèmes développés et des approches proposées, et va apporter à la communauté de recherche en TALN un corpus dédié au domaine éducatif.

Cet article détaille la création et l'utilisation du corpus. Il est structuré comme suit : la deuxième section présente les motivations de la création du corpus. La section 3 présente le corpus. La section 4 décrit un exemple d'utilisation de ce corpus. Enfin, la section 5 présente la conclusion.

2 Motivation de la création du corpus

La disponibilité d'un corpus de discours d'enseignement transcrit et segmenté manuellement offre plusieurs possibilités applicatives. Nous citons dans cette section les motivations qui nous ont conduit

à construire le corpus présenté dans cet article.

2.1 Adaptation automatique de modèles de langage

Le modèle de langage (ML) est l'un des constituants d'un système de reconnaissance de la parole. Un ML a pour but de guider le décodeur à choisir la séquence de mots la plus probable. Il est généralement formé à partir d'une large quantité de données représentatives de la tâche pour laquelle le modèle sera utilisé. Cependant, ce ML n'est pas fiable lorsqu'il s'agit de transcrire des documents oraux traitants d'une autre tâche. Les systèmes dont les modèles de langage sont entraînés à partir de données généralistes ne sont pas performants pour transcrire des données liées à des domaines spécifiques. Les données d'apprentissage d'un modèle de langage sont sous forme de séquences de mots collectés à partir d'un discours réel qui sont transcrits manuellement ou semi-automatiquement. Par conséquent, la construction d'un nouveau ML pour chaque domaine est très coûteuse. L'idée classique est d'utiliser un modèle de langage généraliste et de l'adapter aux données du domaine. Dans le cadre de ce travail, où il s'agit de transcrire des cours magistraux portant sur divers domaines, il est donc nécessaire d'avoir recours à des techniques d'auto-adaptation de ces modèles. L'une de ces techniques, présentée dans cet article, sera évaluée grâce au corpus que nous présentons.

2.2 Segmentation et enrichissement de contenu pédagogique

Les transcriptions automatiques peuvent également servir pour la recherche automatique de matériaux pédagogiques et d'informations complémentaires. En travaillant à la structuration automatique du discours du tuteur, par exemple en découpant ce discours en segments thématiques, il est possible d'extraire de ces segments des mots clés, ou de caractériser ces segments par d'autres moyens, de manière telle qu'il soit possible de lier un segment thématique à un ensemble de sources d'informations disponibles par ailleurs et non produites par le tuteur. Dans un contexte de cours magistral, en exploitant si possible les sources d'informations a minima fournies par l'enseignant, il s'agira de segmenter à la volée, c'est-à-dire en direct au fur et à mesure de la transcription automatique en temps réel, les sorties d'un système de reconnaissance automatique de la parole appliqué sur le discours de l'enseignant. Cette segmentation sera de type thématique : il s'agira de détecter les frontières de zones homogènes au niveau du contenu, qu'il faudra caractériser afin de le lier avec des documents pédagogiques disponibles dans une base de connaissances extérieure (documents extérieurs, issus de bases de données spécifiques, d'encyclopédies en ligne comme Wikipédia, ou d'autres sources du web). Dans la littérature, la tâche de la segmentation thématique est relativement bien explorée en TALN, à travers par exemple les travaux de (Bouchekif *et al.*, 2015), (Galley *et al.*, 2003), (Caillet *et al.*, 2004). Dans ces travaux, la segmentation thématique a généralement été réalisée sur des documents finalisés, et non en temps réel. La réalisation de cette tâche, en particulier pour une approche par apprentissage automatique, nécessite un corpus segmenté thématiquement.

2.3 Obtention d'un découpage de séquences autonomes

Cette tâche vise à découper le cours en une suite de séquences de quelques minutes homogènes, de manière chaque séquence puisse être considérée comme un chapitre du cours. Il s'agit d'être capable de reconstruire le plan du cours. Un niveau de granularité plus fin est également pris en compte, qui consiste à segmenter le cours en zone de description de concepts, dont les frontières peuvent être

délimitées par la synchronisation avec une ou plusieurs diapositives dans le cas de cours magistraux qui en utilisent. Le corpus PASTEL est ainsi annoté en plusieurs niveaux de granularité.

3 Corpus

Cette section présente les données, les conventions d'annotation ainsi que les statistiques du corpus.

3.1 Origine du corpus

Les données utilisées sont issus du projet CominOpenCourseware (COCo) [1] qui met à disposition un certain nombre de vidéos avec des ressources potentielles. Les vidéos collectées concernent des cours magistraux de niveau licence. Les vidéos sont alignées temporellement avec les diapositives de la présentation.

Ainsi quelques vidéos sont issues de la plateforme Umotion [2] : plateforme vidéo de l'Université du Mans.

3.2 Annotations

Il est très important pour la qualité et la fiabilité des annotations que des conventions soit mises en place pour guider les annotateurs humains.

Transcription

Pour la partie de la transcription, le logiciel Transcriber [3] a été utilisé. Transcriber est un outil d'aide à l'annotation manuelle de signal vocal (Barras *et al.*, 2001). Il fournit une interface utilisateur graphique conviviale pour segmenter les enregistrements vocaux de longue durée, les transcrire, et marquer les tours de parole, les changements de sujets et les conditions acoustiques. Les conventions généralement utilisées pour les transcriptions de campagnes d'évaluation (Gravier *et al.*, 2004) ont servi de guide pour transcrire les cours enregistrés. Pour accélérer le traitement, la transcription de référence du corpus PASTEL a été réalisée de manière semi-automatique. Une première transcription est générée par un système de reconnaissance de la parole générique (dont le modèle de langage ne contient pas des données du domaine) avant d'être corrigé par l'annotateur.

Segmentation

Afin de guider la segmentation thématique, nous devons répondre à la question suivante : " Qu'est-ce qu'un sous-thème ? " dans un cours qui est monothématique (l'objet principal du cours). Nous avons décidé de répondre à cette question en fonction des motivations que nous avons énumérées dans la section précédente. Nous sommes partis de l'hypothèse qu'une frontière thématique ne peut se situer qu'au voisinage d'un changement de diapositive pendant le cours. Par conséquent, à chaque changement de diapositive, il est nécessaire d'annoter :

1. s'il y a un changement thématique ou non,
2. l'instant exact du changement thématique défini comme étant positionné entre deux mots,

1. http ://www.comin-ocw.org/
2. umotion.univ-lemans.fr/
3. http ://trans.sourceforge.net/en/presentation.php

3. la granularité du changement thématique (1 ou 2).

La granularité 1 est utilisée pour marquer qu'une nouvelle notion est abordée tout en restant dans le même sous-thème. La granularité 2 est utilisée lorsqu'il y a un changement de sous-thème plus général qui permet d'arrêter l'apprentissage à ce moment-là et de reprendre plus tard l'apprentissage d'autres notions. La granularité 2 permet ainsi de chapitrer le cours, chaque chapitre étant constitué de segments de type 1.

Nous avons aussi ajouté à l'annotation la notion d'interruption : il s'agit de localiser les portions de discours qui correspondent à des moments de gestion de l'attention du public, de problèmes techniques (gérer un problème de vidéo-projecteur par exemple), etc.

L'annotation a été effectuée par deux annotateurs étudiants en master en linguistique à l'aide de l'outil ELAN [4] (Auer *et al.*, 2010). ELAN (**E**UDICO **L**inguistic **A**nnotator) est un outil d'annotation qui permet de créer, éditer, visualiser et rechercher des annotations pour des données vidéo et audio. Chaque annotateur a annoté le corpus conformément aux directives précisées dans la section précédente. Par la suite, les deux annotateurs ont confronté leurs annotations pour trouver un consensus et proposer une annotation finale.

3.3 Statistiques du corpus

Le corpus annoté comprend neuf cours. La durée totale du corpus est d'environ 9 heures. Le tableau 1 illustre quelques statistiques de notre corpus. La deuxième, la troisième et la quatrième colonnes du tableau représentent respectivement le nombre de segments de granularité 1, le nombre de segments de granularité 2 et le nombre de segments de type "interruption". La dernière colonne contient la durée de chaque cours.

Nom du cours	**Gran. 1**	**Gran. 2**	**Interp**	**Durée**
Introduction à l'informatique	31	2	2	1h 04mn 42s
Introduction à l'algorithmique	38	10	3	1h 17mn 28s
Les fonctions dans l'algorithmique	35	3	3	1h 14mn 29s
Réseau sociaux et graphes	43	7	7	1h 05mn 51s
Algoritmique distribuée	72	5	3	1h 16mn 30s
Langage naturel	52	5	5	1h 09mn 35s
Architecture république	49	7	0	1h 21mn 14s
Méthode traditionnelle	12	7	1	0h 41mn 02s
Imagerie	57	0	1	1h 08mn 14s

TABLE 1 – Statistiques du corpus (Interp : Interruption, Gran : Granularité)

4 Adaptation du modèle de langage

Cette section présente les premiers travaux menés dans le projet PASTEL pour l'adaptation du modèle de langage.

4. https ://tla.mpi.nl/tools/tla-tools/elan/

4.1 Expérimentation

Les expériences effectuées pour l'adaptation du modèle de langage s'inspirent du travail présenté dans (Lecorvé *et al.*, 2008). Le processus d'adaptation que nous avons effectué dans ce travail consiste à :

1. Identifier les mots-clés en utilisant le critère TF-IDF à partir des diapositives du cours. Cette étape est précédée par un prétraitement qui vise à lemmatiser le texte en utilisant l'outil MACAON (Nasr *et al.*, 2010) et à supprimer les mots-vides (en anglais *stop word*).
2. Les mots fréquents extraits par TF-IDF vont servir à formuler des requêtes : nous avons pris les 10 mots-clés ayant les scores TF-IDF les plus élevés puis ces mots clés ont été combinés pour effectuer des requêtes constituées chacune de 3 mots.
3. Extraire des documents à partir du web : les requêtes sont soumises à un moteur de recherche sur web (Google) et les pages pointées par les liens retournés sont téléchargées.
4. Nettoyer les données issues du web : les ML ont besoin d'être entraînés sur des corpus propres et normalisés afin de garantir un certain niveau de qualité. Dans notre cas où la source des données d'adaptation est le web, il est important de nettoyer le texte en éliminant les ponctuations, les URLs, les adresses mail et en transformant les formules mathématiques ainsi que les chiffres en des séquences de mots.
5. Construire un modèle de langage du domaine en utilisant les données collectées du web.
6. Adapter le modèle de langage par interpolation linéaire de deux modèles : le modèle générique et le modèle du domaine.

4.2 Résultats

La qualité de l'adaptation des modèles de langage est évaluée à l'aide de la métrique d'évaluation taux d'erreurs mots WER (Pallett, 2003), couramment utilisée dans la littérature pour l'analyse des performances des systèmes de reconnaissance de la parole. Elle se calcule ainsi :

$$WER = \frac{S + I + D}{N} \tag{1}$$

où S est le nombre de mots remplacés par le système , I est le nombre de mots insérés par le système, D est le nombre de mots supprimés par le système, et N est le nombre total de mots dans la référence.

Le tableau 1 présente les premiers résultats expérimentaux. Nous avons adapté le modèle de langage pour les deux cours *Introduction à l'informatique* et *Introduction à l'algorithmique*.

Cours	**% WER sans adaptation**	**% WER avec adaptation**
Introduction à l'informatique	16,2	15,2
Introduction à l'algorithmique	19,6	18,7

TABLE 2 – Résultats d'adaptation du modèle de langage

Les premiers résultats d'adaptation des modèles de langage montrent un gain en terme de WER et sont encourageants pour améliorer la qualité des transcriptions générées pendant les cours.

5 Conclusion

La construction de corpus est une étape cruciale pour tout système de traitement automatique du langage naturel. Nous avons présenté dans cet article un travail portant sur la création d'un corpus annoté, recueilli pour développer une application pour la transcription, la segmentation, l'enrichissement automatique en temps réel et d'autres applications à vocation pédagogique. Ce corpus a été créé dans le cadre du projet PASTEL. Il comprend plus de neuf heures de parole transcrites et annotées. Des expériences préliminaires ont été réalisées pour évaluer notre système de reconnaissance automatique de la parole.

Remerciements

Nous remercions l'agence ANR pour son financement à travers le projet PASTEL sous le numéro de contrat ANR-16-CE33-0007.

Références

AUER E., RUSSEL A., SLOETJES H., WITTENBURG P., SCHREER O., MASNIERI S., SCHNEIDER D. & TSCHÖPEL S. (2010). Elan as flexible annotation framework for sound and image processing detectors. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, p. 890–893 : European Language Resources Association (ELRA).

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1-2), 5–22.

BOUCHEKIF A., CHARLET G. D. N. C. D. & ESTÈVE Y. (2015). Segmentation et titrage automatique de journaux télévisés. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

CAILLET M., PESSIOT J.-F., AMINI M.-R. & GALLINARI P. (2004). Unsupervised learning with term clustering for thematic segmentation of texts. In *Coupling approaches, coupling media and coupling languages for information retrieval*, p. 648–657 : LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

GALLEY M., MCKEOWN K. R., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

GRAVIER G., BONASTRE J., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. *Proc. Journées d'Etude sur la Parole (JEP)*.

HO I., KIYOHARA H., SUGIMOTO A. & YANA K. (2005). Enhancing global and synchronous distance learning and teaching by using instant transcript and translation. In *Cyberworlds, 2005. International Conference on*, p. 5–pp : IEEE.

HWANG W.-Y., SHADIEV R., KUO T. C. & CHEN N.-S. (2012). Effects of speech-to-text recognition application on learning performance in synchronous cyber classrooms. *Journal of Educational Technology & Society*, **15**(1), 367.

LECORVÉ G., GRAVIER G. & SÉBILLOT P. (2008). An unsupervised web-based topic language model adaptation method. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, p. 5081–5084 : IEEE.

NASR A., BÉCHET F. & REY J.-F. (2010). Macaon : Une chaîne linguistique pour le traitement de graphes de mots. In *Traitement Automatique des Langues Naturelles*.

PALLETT D. S. (2003). A look at nist's benchmark asr tests : past, present, and future. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, p. 483–488 : IEEE.

SHADIEV R., HWANG W.-Y., CHEN N.-S. & YUEH-MIN H. (2014). Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society*, **17**(4), 65.