Alignement multimodal de ressources éducatives et scientifiques

Hugo Mougard
Université de Nantes
hugo.mougard@univ-nantes.fr

Résumé. Cet article présente certaines questions de recherche liées au projet COCo [1]. L'ambition de ce projet est de valoriser les ressources éducatives et académiques en exploitant au mieux les différents médias disponibles (vidéos de cours ou de présentations d'articles, manuels éducatifs, articles scientifiques, présentations, etc). Dans un premier temps, nous décrirons le problème d'utilisation jointe de ressources multimédias éducatives ou scientifiques pour ensuite introduire l'état de l'art dans les domaines concernés. Cela nous permettra de présenter quelques questions de recherche sur lesquelles porteront des études ultérieures. Enfin nous finirons en introduisant trois prototypes développés pour analyser ces questions.

Abstract.

Multimodal Alignment of Educative and Scientific Resources
This article presents some questions linked to the COCo [1] project. The ambition of this project is to better exploit educative and academic resources by leveraging the different available supports (courses and conference talks videos, scientific articles, textbooks, slides, etc). We will first describe the problem of joint educative or scientific multimedia resources use. We will then give an overview of the related state of the art that will allow us to introduce a few research questions that will be the subject of further studies. Finally we will introduce three research prototypes that are helpful to start to investigate those research questions.

Mots-clés : alignement, e-research, e-learning, multimodalité.

Keywords: alignment, e-research, e-learning, multimodality.

1 Introduction

À chaque avancée technologique, les professionnels de l'éducation doivent repenser et adapter leurs méthodes et matériels pour tirer partie des nouvelles possibilités. La récente arrivée des formations en ligne ouvertes à tous (*massive open online courses* ou *MOOC*s en anglais) a ainsi entraîné de nombreux changements d'usage pour les institutions éducatives et les étudiants. Un des points essentiels ayant permis l'essor des *MOOC*s est la disponibilité de plus en plus fréquente d'excellentes connexions internet, rendant les cours vidéos téléchargeables rapidement. Malgré ce rôle central des vidéos dans les nouveaux cours disponibles sur internet, celles-ci restent souvent traitées comme des blocs indépendants et ne sont pas coordonnées au mieux avec le reste du matériel éducatif. Ce constat constitue le point de départ des études détaillées par la suite, qui ont pour but de contribuer au changement de cet état de fait.

De plus, la disponibilité croissante de matériel vidéo ne se limite pas aux plates-formes de *MOOC*s mais concerne aussi la recherche. Ce phénomène s'explique par plusieurs facteurs : (1) le travail fondateur de VIDEOLECTURES [2], qui permet d'accéder à de nombreuses conférences et tutoriels scientifiques filmés, (2) les coûts de production réduits grâce à des initiatives comme MATTERHORN [3], (3) l'intérêt grandissant des institutions pour promouvoir leurs événements, et (4) le mouvement grandissant en faveur des données libres (open data). Ce phénomène d'augmentation de la quantité de matériel de recherche multimédia disponible se constate en particulier pour les vidéos correspondant aux présentations faites lors de conférences. La disponibilité de ces vidéos représente une opportunité sans précédent car elles sont souvent accompagnées de l'article qui leur est associé dans les actes de la conférence, ce qui constitue un cadre idéal pour expérimenter dans le domaine de l'intégration profonde de différents médias (hypermédia). En effet, traditionnellement,

1. http://www.comin-ocw.org/
2. http://videolectures.net
3. http://opencast.org/matterhorn/

les résultats scientifiques sont étudiés au travers de l'article dans lequel ils sont exposés, et peu de médias alternatifs sont disponibles pour le compléter. Pour guider la recherche qui suit, nous considérons que cette étude de l'article seule n'est pas optimale et qu'il faut inventer de nouveaux moyens de combiner les différents supports disponibles pour fournir une expérience riche et efficace aux utilisateurs de systèmes éducatifs ou de recherche.

2 Description du problème

Dans ce travail au sein du projet COCo, nous visons l'exploitation pleine des ressources multimédias dans un cadre éducatif ou scientifique. Pour atteindre ce potentiel nous nous appuyons sur le travail des communautés du traitement de la langue mais aussi du web sémantique, de l'hypertexte, de l'hypermédia et du multimédia.

Le système cible que nous souhaitons obtenir est composé de deux sous-systèmes qui interagissent :

1. Le premier système prend en entrée un ensemble de documents multimédia, par exemple un article scientifique au format PDF et un enregistrement de sa présentation donnée durant une conférence, et renvoie des liens typés allant d'un document à l'autre, avec des types allant de la thématique à la rhétorique (*e.g.* une définition dans un article pourra être liée au segment vidéo expliquant le concept défini à l'aide d'un lien de type `explication`. Dans le sens inverse, le segment vidéo pourra être lié à la définition par un lien de type `définition formelle`).
2. Le second système utilise ces liens pour proposer à l'utilisateur une navigation intuitive et précise dans le matériel pédagogique. Il doit aussi pouvoir retransmettre un retour utilisateur (implicite ou explicite) afin que le premier système puisse s'améliorer au fur à mesure de son utilisation.

Le but premier du système est de transformer les vidéos en hypervidéos, selon le modèle de Sadallah *et al.* (2012) et les articles en hyperarticles selon un modèle similaire. Une vidéo peut n'être utilisée que comme un bloc atomique, séquentiel, pour éclaircir un concept (Navarrete & Blat, 2002) ou au contraire être intégrée en profondeur dans d'autres ressources pour permettre une navigation aisée et non linéaire du matériel, elle devient alors une hypervidéo (Sadallah *et al.*, 2012).

Les types d'hyperliens (liens inter-documents) à utiliser et les possibilités de navigations offertes doivent tenir compte des caractéristiques complémentaires des principaux média disponibles. Nous les détaillons ci-dessous avec l'exemple de l'article scientifique et de sa présentation vidéo :

Capacité à être lu en diagonale Une présentation enregistrée est dure à parcourir et en particulier à survoler. Cette difficulté peut être contournée par un alignement avec son article grâce à la structure de ce dernier.

Définitions formelles Quand un utilisateur souhaite "se pencher sur les maths", une vidéo de présentation n'est habituellement pas suffisante. L'article peut être utilisé pour compléter efficacement certaines parties de la présentation avec des définition précises.

Références Les références présentes dans un article peuvent soit être utilisées directement soit constituer des pistes intéressantes pour des hyperliens. Il est par exemple possible d'inclure un morceau d'une présentation vidéo d'un article référencé pour éclaircir un point compliqué dans la présentation qui intéresse l'utilisateur, ou plus précisément une partie précise de celle-ci qui répond au besoin d'information de l'utilisateur.

Détails Le manque de détails dans une présentation vidéo est double : (1) Le manque de détails dans le contenu d'une présentation induit par les contraintes de temps imposées durant les conférences est solvable d'une manière similaire au manque de définitions formelles par un alignement avec l'article correspondant, et (2) Les parties de l'article correspondant à la présentation vidéo qui n'ont pas été utilisées peuvent aussi être intéressantes pour l'utilisateur. Un alignement permet de les récupérer en considérant les parties non alignées entre les deux média.

Figures La nature dynamique d'une présentation vidéo lui permet souvent de mieux expliquer les figures complexes qu'un article. Intégrer un clip de l'auteur qui explique certaines parties des concepts cruciaux de ses figures peut faire gagner un temps précieux à l'utilisateur.

Biais de l'auteur Les vidéos de présentations sont habituellement plus subjectives que les articles correspondants. Il est en conséquence plus facile d'y discerner le biais de l'auteur par rapport à son travail.

Vision globale Les présentations ont pour but principal de faire passer l'idée forte d'une recherche scientifique dans le contexte qui la met le mieux en valeur. Les articles correspondant aux présentations ont pour objectif de détailler davantage la recherche scientifique exposée, rendant la vision globale du contexte plus délicate à acquérir. Encore une fois combiner les deux média est donc profitable.

Synthèse D'un côté, le texte de l'article est plus aisé à résumer que la transcription d'une vidéo, de l'autre, le matériel est habituellement plus synthétique dans une présentation que dans un article scientifique. Conséquemment, les deux médias peuvent contribuer à améliorer la capacité de synthèse du système.

Une fois les types d'hyperliens formalisés pour satisfaire aux différents besoins de navigation soulignés par ces complémentarités, il est nécessaire de concevoir une interface de navigation pour les exploiter.

Nous prévoyons deux versions de ce système, l'une orientée e-learning sur la plate-forme COCo, l'autre orientée e-research sur la plate-forme VIDEOLECTURES :

E-learning sur COCo Le système s'intègrera ici dans une plate-forme de diffusion de ressources éducatives. Son but sera de dépasser le paradigme de la vidéo reine qui est dominant dans les plates-formes actuelles. Il devra ainsi permettre une navigation efficace entre le cours principal, les manuels et les lectures complémentaires.

E-research sur VIDEOLECTURES Sur VIDEOLECTURES, l'objectif est principalement d'utiliser les nombreux couples d'article et de leur présentation vidéo disponibles afin d'améliorer l'expérience d'étude de résultat scientifique.

3 État de l'art

Le système présenté ci-dessus nécessite des algorithmes et idées de plusieurs domaines de l'informatique. Dans cet article, nous nous concentrons sur trois d'entre eux :

1. l'alignement multimodal ;
2. les typologies spécialisées pour le traitement informatique des ressources scientifiques et éducatives ;
3. l'évaluation extrinsèque et l'apprentissage par renforcement, qui sont nécessaires pour pallier le manque de données et qui sont envisagés de manière jointe.

Ces trois domaines ne couvrent pas le problème de manière exhaustive mais fournissent déjà un point de départ satisfaisant pour notre travail.

3.1 Alignement multimodal

La création des liens entre les différents médias est au cœur d'un système hypermédia. Nous voyons ces liens comme la concrétisation d'un alignement multimodal. Brunning (2010) distingue trois granularités d'alignement : (1) niveau du document : identifier des paires de documents semblables au sein de corpus comparables ou parallèles, (2) niveau de la phrase : identifier les phrases similaires dans deux documents comparables ou parallèles, et (3) niveau du mot : trouver des mots similaires dans deux phrases comparables ou similaires. De par la nature de notre corpus (paires d'articles et leur présentation), l'alignement au niveau du document n'est pas nécessaire. Par ailleurs, l'alignement des paires au niveau du mot peut servir pour certaines tâches (lien d'entités) mais n'est pas nécessaire à la majorité des cas d'utilisations mentionnés en partie 2. Cet état de l'art se concentrera donc sur l'alignement au niveau de la phrase.

En outre, l'alignement multimodal est un sujet peu traité de manière directe. Nous le verrons donc comme la combinaison de trois facteurs :

1. des prétraitements et adaptions spécifiques aux extractions textuelles des modalités traitées ;
2. une mesure de similarité entre les différentes unités choisies (*e.g.* phrases pour un article scientifique, fenêtres fixes, groupes de souffle ou phrases reconstituées pour une vidéo) ;
3. un alignement entre les différents modes utilisant cette mesure.

Prétraitements et adaptions spécifiques aux extractions textuelles Le caractère multimodal des documents traités requiert une attention particulière. Dans cet article, Nous exploitons les matériels multimodaux par l'intermédiaire de l'extraction de leur modalité textuelle et le résultat de cette extraction est bruité (la modalité textuelle correspond au texte extrait du pdf, pour les documents écrits, et à la transcription automatique de la parole, pour les documents vidéos). Pour comprendre l'impact de ce bruit sur une chaîne de traitements classique, il suffit de remarquer qu'une des premières étapes nécessaires au traitement de la langue — la séparation du texte d'entrée en phrases — n'est plus possible aux

précisions habituelles sur du texte non bruité (98%) ; les erreurs de cette étape de pré-traîtement ont des répercussions multiplicatives sur les erreurs de la chaîne complète. En effet, la ponctuation ne peut pas être inférée à partir du seul son de la vidéo et n'est donc renvoyée par aucun système de transcription (hormis au prix d'étapes de post-traitement de la sortie du système). De même, alors que l'on peut détecter automatiquement des segments porteurs de sens dans un texte (paragraphes, sections), cette détection est beaucoup plus compliquée dans une extraction. À ces problèmes de structure manquante s'ajoutent les 20% de taux d'erreur sur les mots transcrits malgré l'utilisation du système état de l'art de TRANSLECTURES.

Pour ces raisons, depuis 2001, des campagnes d'évaluation sont organisées pour améliorer le traitement des supports multimédias (Smeaton *et al.*, 2001). Plus récemment, une initiative spécialisée a vu le jour sous le nom de MEDIAEVAL. Une de ses tâches, SEARCH & ANCHORING IN VIDEO ARCHIVES [4], évalue des systèmes sur un cas d'utilisation très courant : un utilisateur recherche par une requête textuelle une information dans un corpus vidéo et suit les liens proposés par le système depuis les résultats fournis pour satisfaire son besoin d'information. Ce cas d'utilisation est très proche de ceux rencontrés dans une application hypermédia, où un utilisateur suit des hyperliens pour satisfaire son besoin d'information. Cette tâche est donc d'un grand intérêt pour sa similarité aux besoins de notre étude. Des articles issus de ces campagnes d'évaluation, on peut relever des approches d'apprentissage automatique pour segmenter les transcriptions (Galuščáková, 2013) afin de maximiser les chances de renvoyer un segment vidéo qui aura du sens pour un utilisateur ou encore des travaux d'adapation des mesures de recherche d'information classiques aux supports multimédias (Eskevich *et al.*, 2012).

Mesures de similarité Parmi les différentes méthodes que l'on trouve dans la littérature pour comparer deux chaînes de caractères, deux grandes classes émergent : (1) les méthodes basées sur le calcul d'un coût de l'alignement optimal des deux chaînes, et (2) les méthodes basées sur un apprentissage supervisé.

L'étude des chaînes de caractères (*Stringology*) remonte aux années 1960 et est le domaine dans lequel ont été publiées la majorités des approches par alignement. La première distance qui utilise ce concept est due à Hamming (1950). Elle correspond au le nombre d'erreurs dans l'alignement direct, caractère par caractère, de deux séquences de caractères. Levenshtein (1966) a introduit par la suite une distance prenant en compte le nombre pondéré de suppressions, insertions et substitutions nécessaires pour transformer la chaîne source en la chaîne cible. L'adaptation de cette distance à des domaines de spécialité (*e.g.* traitements en temps réel, biologie, etc) est encore l'objet d'articles scientifiques (Uhl & Wild, 2010). On peut citer certaines variantes de la distance de Levenshtein comme la distance de Jaro-Winkler (Winkler, 1990), plus adaptée aux chaînes courtes, ou la distance de Needleman & Wunsch, qui, parmi d'autres propriétés détaillées dans la section suivante, permet de donner un score de similarité plus précis aux couples d'objets alignés que les trois poids des opérations de la distance de Levenshtein et permet ainsi d'obtenir une distance plus fine.

Outre les mesures de similarité issues du domaine de l'étude des chaînes de caractères, de nombreuses approches par apprentissage permettent de rendre compte de la similarité de deux phrases. Par exemple, Hatzivassiloglou *et al.* (2001) ont créé SIMFINDER, un module de résumé multi-documents qui regroupe des phrases de même sens provenant de différents documents fournis, afin de sélectionner les phrases à inclure dans le résumé. Ils utilisent une approche d'apprentissage supervisé avec divers traits comprenant les groupes nominaux, noms propres, sens dans WordNet et les comptes de mots. Ou encore, dans le domaine de l'identification de paraphrases, Madnani *et al.* (2012) détectent des paraphrases en se basant sur des mesures de traduction automatique et Socher *et al.* (2011) les détectent au moyen d'un réseau de neurones profond. Smith *et al.* (2010), pour leur part, utilisent les alignements de mots utilisés traditionnellement en traduction automatique ainsi que des traits complémentaires pour parvenir aux mêmes fins. Nelken & Shieber (2006), quant à eux, calculent une régression logistique sur la similarité cosinus des TF-IDF d'une paire de phrases afin de déterminer si elle est ou non paraphrastique. Dans le cadre multilingue, Munteanu & Marcu (2005) utilisent l'alignement au niveau des mots entre deux phrases pour déterminer si elles sont paraphrastiques.

Alignement de phrases dans deux documents comparables De la même manière que pour la définition d'une mesure de similarité, la stratégie globale d'alignement à utiliser pour aligner deux séquences d'objets dépend fortement du domaine d'application et de contraintes spécifiques. En conséquence, les algorithmes ont également été proposés dans des domaines variés.

En bio-informatique, un besoin courant est d'aligner deux génomes ou deux séquences d'acides aminés. Ces alignements sont contraints : l'ADN a des séquences non informatives qu'il ne faut pas considérer, les séquences d'acides aminés peuvent varier sans grands effets biologiques (certains acides aminés sont très proches dans leur fonction) et peuvent

4. http://www.multimediaeval.org/mediaeval2015/searchandanchor2015/

également contenir des séquences non informatives. Ces contraintes ont donné naissance à de nouveaux algorithmes. Comme présenté dans l'article de Navarro (2001), l'algorithme central qui a répondu à la contrainte — l'algorithme de Needleman-Wunsch (Needleman & Wunsch, 1970) — a aussi été inventé dans dans la communauté du traitement automatique de la parole sous le nom de Dynamic Time Warping (Vintsyuk, 1968). Cet algorithme a été adapté par Smith & Waterman (1981) pour accentuer l'importance des alignements locaux et diminuer l'importance des trous (*gaps*). Dans leur article, Nelken & Shieber (2006) adaptent la distance de Needleman & Wunsch pour prendre en compte les besoins spécifiques de l'alignement de texte : ne pas pénaliser les alignements de n objets vers 1 objet en particulier. (Bott & Saggion, 2011) proposent, quant à eux, une méthode qui s'appuie sur des modèles de Markov cachés (*Hidden Markov Models* ou *HMM*s) pour modéliser l'alignement de textes contraints (le texte cible est la simplification du texte source dans leur cas).

Dans la continuation des algorithmes basés sur la distance de Levenshtein, on trouve l'algorithme de Myers & Miller (1988), qui calcule en espace linéaire l'alignement entre deux documents. Cet algorithme est la base de tous les outils modernes de différenciation de documents (*diff tools*) tels que ceux trouvés dans les gestionnaires de version comme GIT [5] ou SUBVERSION [6].

Cette famille d'algorithmes est très puissante pour modéliser certains alignements mais manque cependant de généralité : en particulier, un document et sa réorganisation par blocs (*e.g.* un article scientifique et sa présentation orale qui réarrangerait les sections pour mieux mettre en valeur certains points) rompent la linéarité nécessaire au bon fonctionnement de ces algorithmes — les opérations d'insertion, substitution et suppression ne permettent pas de modéliser l'alignement de telles paires de documents sans multiplier les non-correspondances (*mismatchs*). C'est pour cela que des algorithmes différents ont été proposés dans d'autres domaines où ces lacunes sont problématiques.

Par exemple, pour évaluer les systèmes de traduction automatique, Snover *et al.* (2009) introduisent TER-Plus, une extension de la distance de Levenshtein qui prend en compte les réorganisations par blocs, au prix de la NP-complétude du calcul. Dans un domaine différent, Chen *et al.* (2009) modélisent la structure d'un document au sein d'un corpus donné par un processus Bayésien appliqué aux thèmes qu'il aborde. L'alignement de deux documents devient alors la comparaison de l'issue de ce processus pour ces deux documents. Un autre moyen utilisé pour aboutir à des concepts d'alignement plus expressifs a été introduit par Barzilay & Elhadad (2003). Il s'agit d'abord de regrouper les objets dans des ensembles homogènes (*clusters*) sur un critère donné et de manière jointe dans les deux documents à aligner. Une fois ces *clusters* formés, un alignement local peut être effectué au moyen d'un des algorithmes de la famille citée ci-dessus. Ainsi, le réorganisation par blocs des *clusters* peut être modélisée sans problème durant leur formation, hors du processus d'alignement et permettant à celui-ci d'utiliser les algorithmes performants dérivés de l'algorithme de Levenshtein.

3.2 Typologies

L'intérêt de l'expérience d'apprentissage ou d'étude proposée à l'utilisateur dépend directement de la qualité des typologies de liens utilisées pour lier les documents. Pour correctement couvrir les cas d'utilisations envisagés, nous distinguons principalement deux types de typologies : (1) les typologies à ambition ontologique, qui cherchent à modéliser le savoir scientifique ou éducatif présenté, et (2) les typologies rhétoriques, nécessaires dès lors que l'on constate que les communications scientifiques sont éminemment rhétoriques par nature (Bazerman *et al.*, 1988).

Ces typologies ont été spécifiquement étudiées depuis plus de quinze ans (Teufel *et al.*, 1999) et ont abouti à plusieurs schémas d'annotation (Guo *et al.*, 2010). Parmi ceux-ci, nous retenons en particulier le schéma ARGUMENT ZONING II (Teufel *et al.*, 2009) pour la dimension rhétorique qui partitionne un article scientifique en quinze types d'arguments. Par exemple, on y trouve les classes AIM pour souligner un objectif de recherche ou USE pour mentionner l'utilisation des travaux d'autrui dans un travail, etc. Nous retenons aussi le schéma CORESC (Liakata *et al.*, 2010) qui adopte une approche ontologique et a été utilisé en conjonction avec ARGUMENT ZONING II. Liakata *et al.* ont montré que les classes des deux schémas d'annotation n'étaient pas redondantes et couvraient ainsi correctement les différents aspects d'une communication scientifique.

5. http://git-scm.com/
6. https://subversion.apache.org/

3.3 Évaluation et apprentissage par renforcement

Mesurer la qualité d'applications hypermédias est difficile pour deux raisons : (1) Cette qualité est subjective, dépend des besoins d'information de l'utilisateur, de ses habitudes d'apprentissage ou d'étude et de son niveau d'intérêt pour le sujet exposé, et (2) Un résultat négatif comme positif de l'évaluation peut être dû à de nombreuses briques différentes du système, étant donnée sa complexité. Il devient alors difficile d'interpréter les résultats pour améliorer le système.

De manière générale, il est possible d'employer deux stratégies complémentaires pour évaluer un système : (1) mesurer la manière dont il accomplit sa tâche caractéristique (*e.g.*, dans notre cas, aligner des documents et extraire des hyperliens de cet alignement), et (2) mesurer le degré auquel il satisfait les utilisateurs ou systèmes qui consomment ses sorties (*e.g.*, dans notre cas, la qualité de l'étude ou de l'apprentissage de l'utilisateur). Le premier point correspond à une évaluation intrinsèque alors que le second point correspond à une évaluation extrinsèque.

L'évaluation intrinsèque nécessite traditionnellement un ensemble de données annotées, formant une vérité terrain (*gold standard*) afin de pouvoir comparer les sorties du système à des sorties correctes. Il est toutefois prohibitif de créer un tel ensemble de données pour un système hypermédia éducatif complet (qui devrait donc comporter des liens corrects sur suffisament de types entre suffisament de documents pour assurer une significativité statistique des expérimentations). Pour envisager ce type d'évaluation il faut donc se concentrer tour à tour sur des sous-parties du système complet. Nous envisageons un découpage en trois sous-systèmes évaluables :

1. l'alignement, sur le corpus Britannica qui est classique dans ce domaine (Barzilay & Elhadad, 2003; Nelken & Shieber, 2006). Il consiste en plusieurs articles déclinés en deux versions : une version tirée de l'encylopedie Britannica standard et l'autre d'une version simplifiée rédigée de manière indépendante ;
2. la multimodalité, sur la tâche SEARCH & ANCHORING IN VIDEO ARCHIVES de la campagne MEDIAEVAL. Cette campagne propose une référence annotée par des humains pour évaluer la qualité d'un système de recherche d'information hypermédia ;
3. la création d'hyperliens, au cours de la même tâche de MEDIAEVAL.

Dans un premier temps, ces évaluations permettront de se situer par rapport aux approches de la littérature et d'avoir des retours rapides durant les phases de développement de ces sous-systèmes. Elles devront cependant être complétées par une évaluation plus complète, extrinsèque. Les évaluations extrinsèques sont souvent menées par test A/B : les utilisateurs se voient proposer aléatoirement un système témoin ou un système à évaluer et la comparaison des deux repose sur l'analyse de leurs comportements. Dans ce cadre, Radlinski *et al.* (2008) montrent que les métriques facilement collectables (*e.g.* clics, temps de visite) ne reflètent pas la qualité des systèmes évalués. Pour pallier ce problème, ils introduisent un mécanisme astucieux de combinaison des systèmes qui permet une évaluation correcte des performances.

La mise en place d'un dispositif analogue d'utilisation des retours des utilisateurs est nécessaire pour apprendre par renforcement (ce qui, comme mentionné en section introduction de cet état de l'art, est nécessaire pour pallier le manque de données). Radlinski *et al.* (2008) montrent qu'en conséquence cela peut être fait conjointement. Le processus de développement du système consiste alors à présenter aux utilisateurs un système évalué intrinsèquement puis à l'évaluer de manière extrinsèque et régulière pour effectuer et mesurer des améliorations.

4 Questions de recherche

Alignement Nous distinguons trois grandes questions :

1. Comment modéliser les alignements qui tolèrent les réorganisations par blocs pour pouvoir les apprendre efficacement ?
2. Comment segmenter le texte extrait des documents écrits et celui provenant de la transcription automatique de la vidéo pour que leur alignement et leur parcours soient aisés ?
3. Quelles sont les mesures de similarité pertinentes dans un contexte multimodal et faut-il leur apporter des modifications pour une pleine efficacité dans la tâche d'alignement au niveau du document ?

Répondre à la question (1) constitue un premier pas vers une adaptation des algorithmes d'alignement de la famille de l'étude des chaînes de caractères aux problèmes nécessitant une tolérance aux réorganisations par blocs. Cette adaptation permettrait d'avoir un cadre théorique et opérationnel fiable pour gérer les alignements multimodaux. Les réponses aux questions (2) et (3), quant à elles, sont primordiales pour pallier le manque de structure des modalités textuelles extraites et sont certainement les questions centrales de la gestion de la multimodalité dans cette étude.

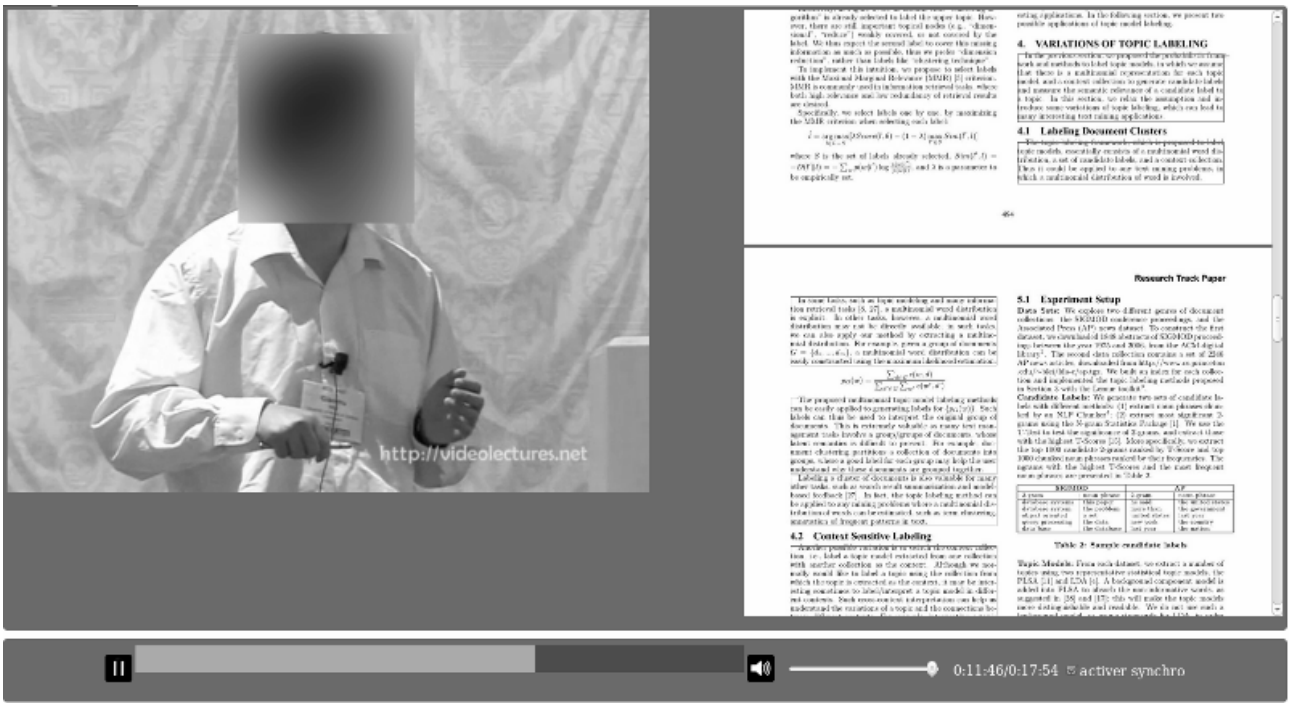

FIGURE 1 – Capture d'écran du prototype de navigation jointe.

Typologie Les schémas d'annotation rhétorique offrent de grands bénéfices aux utilisateurs mais il est extrêmement coûteux de produire des corpus annotés pour entraîner des systèmes d'apprentissage supervisé. La question de recherche intéressante qui découle de ce constat est la suivante : est-il possible d'apprendre à appliquer ces schémas de manière non supervisée ? Des études se sont récemment orientées vers ces questions (Guo *et al.*, 2011) mais utilisent encore des données annotées ou des méthodes d'apprentissage actif.

Évaluation L'approche envisagée est basée sur le travail de Radlinski *et al.* (2008) traitant des systèmes de recommandation. Les applications hypermédias, bien qu'elles aient de nombreuses similarités avec ces systèmes (*i.e.* il est possible de voir les hyperliens comme des recommandations faites aux utilisateurs pour satisfaire leurs besoins d'information), ont des caractéristiques propres. Il est donc nécessaire d'adapter les stratégies et métriques de la littérature pour obtenir un retour utilisateur automatique efficace. Ce travail est en soi une question de recherche à part entière.

5 Prototypes actuels

Nous avons développé trois prototypes [7] pour étudier certaines des questions de recherche mentionnées en Section 4. À ce stade les prototypes se concentrent sur l'alignement multimodal non typé. Le code est disponible sur GITHUB [8] [9]

Illustré par la Figure 1, le premier prototype permet de naviguer de manière jointe dans une présentation enregistrée et dans un article si leur alignement est disponible : cliquer sur un paragraphe entraîne la lecture du segment vidéo lui correspondant et de la même manière, lire la vidéo surligne les parties de l'article liées au sujet courant. Ce prototype utilise POPPLER [10] pour délimiter spatialement les paragraphes d'un article.

Le deuxième prototype, montré en Figure 2, sert à calculer les alignements entre un article scientifique et la vidéo de la présentation correspondante et à étudier leur qualité. Il implémente pour l'instant une approche de base en utilisant les similarités cosinus sur les TF-IDF des phrases de l'article et des segments de la transcription vidéo pour aligner les deux

7. http://alignement.comin-ocw.org/
8. https://github.com/m09/alignment-demo
9. https://github.com/m09/alignment
10. http://poppler.freedesktop.org/

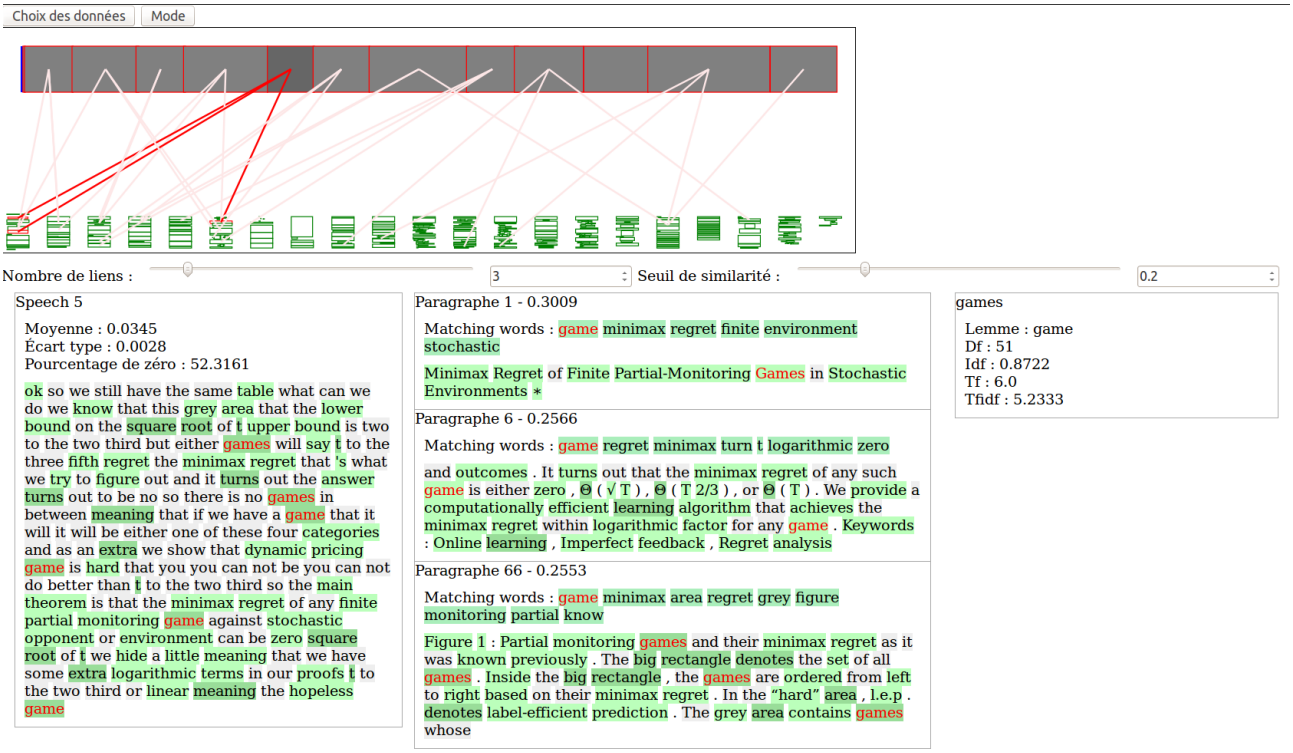

FIGURE 2 – Capture d'écran du prototype d'alignement.

médias. Ces calculs sont réalisés en python avec le framework NLTK [11]. Il présente en résultat les alignements obtenus sur une interface web réalisée avec D3.JS [12].

Quant au troisième prototype, il se concentre sur l'implémentation des algorithmes et métriques de la littérature et l'utilisation des corpus liés (en particulier les approches de Barzilay & Elhadad (2003) et Nelken & Shieber (2006)). Il utilise le framework UIMA [13], les librairies open source OPENNLP [14] et GROBID.

Choix d'utilisabilité Le design d'interface et l'étude d'utilisabilité constituent un travail en cours de réalisation. Pour l'instant, les prototypes sont pensés pour favoriser des itérations de développement rapides et une analyse facile pour les développeurs au détriment du confort d'utilisation pour un utilisateur non expert.

6 Conclusion

Dans cet article, nous avons présenté le problème d'utilisation optimale des ressources éducatives et de recherche en cours d'étude dans le projet COCO. Il mobilise des idées et approches de nombreux domaines de l'informatique. Nous avons ensuite introduit la littérature sur lesquels notre étude s'appuie. Cela nous a permis de discuter quelques pistes pour de futures recherches : (1) la modélisation des alignements qui tolèrent des réorganisations par blocs, (2) les mesures de similarité et segmentations à utiliser pour limiter l'effet détériorant des modalités textuelles bruitées extraites sur le système d'alignement, (3) l'annotation automatique des documents scientifiques respectant des schémas d'annotation rhétoriques et ontologiques et minimisant les données annotées nécessaires, et (4) l'évaluation extrinsèque et l'apprentissage par renforcement dans le cadre d'un système hypermédia. Nous avons fini par présenter trois logiciels en cours de développement qui permettent de débuter l'analyse de ces questions.

11. http://www.nltk.org/
12. http://d3js.org/
13. https://uima.apache.org/uimafit.html
14. https://opennlp.apache.org/

7 Remerciements

Je remercie mes co-auteurs — Matthieu RIOU, Colin DE LA HIGUERA, Solen QUINIOU et Olivier AUBERT — de l'article (Mougard *et al.*, 2015) sur lequel sont basées les sections de cet article qui ne présentent ni l'état de l'art ni les questions de recherche liées au problème étudié.

Nous remercions aussi l'Agence Nationale de la Recherche pour son soutien au programme « Investissements pour le Futur » sous la référence ANR-JO-LABX-07-0J (projet COCO). Nous remercions aussi les collègues de de JSI, Ljubljana pour l'aide qu'ils nous ont apportée ainsi que VIDEOLECTURES pour le matériel multimédia mis à disposition.

References

BARZILAY R. & ELHADAD N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, p. 25–32: Association for Computational Linguistics.

BAZERMAN C. *et al.* (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. University of Wisconsin Press Madison.

BOTT S. & SAGGION H. (2011). An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, p. 20–26, Stroudsburg, PA, USA: Association for Computational Linguistics.

BRUNNING J. J. J. (2010). *Alignment Models and Algorithms for Statistical Machine Translation*. PhD thesis, University of Cambridge.

CHEN H., BRANAVAN S., BARZILAY R., KARGER D. R. *et al.* (2009). Content modeling using latent permutations. *Journal of Artificial Intelligence Research*, **36**(1), 129–163.

ESKEVICH M., MAGDY W. & JONES G. J. (2012). New metrics for meaningful evaluation of informally structured speech retrieval. In *Advances in Information Retrieval*, p. 170–181. Springer.

GALUŠČÁKOVÁ P. (2013). Segmentation strategies for passage retrieval in audio-visual documents. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, p. 1143–1143, New York, NY, USA: ACM.

GUO Y., KORHONEN A., LIAKATA M., KAROLINSKA I. S., SUN L. & STENIUS U. (2010). Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, p. 99–107, Stroudsburg, PA, USA: Association for Computational Linguistics.

GUO Y., KORHONEN A. & POIBEAU T. (2011). A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 273–283, Stroudsburg, PA, USA: Association for Computational Linguistics.

HAMMING R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, **29**(2), 147–160.

HATZIVASSILOGLOU V., KLAVANS J. L., HOLCOMBE M. L., BARZILAY R., YEN KAN M. & MCKEOWN K. R. (2001). Simfinder: A flexible clustering tool for summarization. In *In Proceedings of the NAACL Workshop on Automatic Summarization*, p. 41–49.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, p. 707–710.

LIAKATA M., TEUFEL S., SIDDHARTHAN A. & BATCHELOR C. R. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

MADNANI N., TETREAULT J. & CHODOROW M. (2012). Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, p. 182–190, Stroudsburg, PA, USA: Association for Computational Linguistics.

MOUGARD H., RIOU M., DE LA HIGUERA C., QUINIOU S. & AUBERT O. (2015). The paper or the video: Why choose? In *Proceedings of the Companion Publication of the 24th International Conference on World Wide Web Companion*, WWW Companion '15, p. In press, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

MUNTEANU D. S. & MARCU D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, **31**(4), 477–504.

MYERS E. W. & MILLER W. (1988). Optimal alignments in linear space. *Computer applications in the biosciences: CABIOS*, **4**(1), 11–17.

NAVARRETE T. & BLAT J. (2002). VideoGIS: Segmenting and indexing video based on geographic information. In *5th AGILE Conference on Geographic Information Science*, p. 1–9.

NAVARRO G. (2001). A guided tour to approximate string matching. *ACM Comput. Surv.*, **33**(1), 31–88.

NEEDLEMAN S. B. & WUNSCH C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3), 443–453.

NELKEN R. & SHIEBER S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *In Proc. EACL*: Association for Computational Linguistics.

RADLINSKI F., KURUP M. & JOACHIMS T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, p. 43–52: ACM.

SADALLAH M., AUBERT O. & PRIÉ Y. (2012). CHM: an Annotation- and Component-based Hypervideo Model for the Web. *Multimedia Tools and Applications*.

SMEATON A. F., OVER P. & TABAN R. (2001). The TREC-2001 video track report. *Proceedings of TREC-2001*.

SMITH J. R., QUIRK C. & TOUTANOVA K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 403–411: Association for Computational Linguistics.

SMITH T. F. & WATERMAN M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1), 195–197.

SNOVER M., MADNANI N., DORR B. & SCHWARTZ R. (2009). Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, **23**(2-3), 117–127.

SOCHER R., HUANG E. H., PENNIN J., MANNING C. D. & NG A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, p. 801–809.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, p. 110–117, Stroudsburg, PA, USA: Association for Computational Linguistics.

TEUFEL S., SIDDHARTHAN A. & BATCHELOR C. (2009). Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, p. 1493–1502: Association for Computational Linguistics.

UHL A. & WILD P. (2010). Enhancing iris matching using levenshtein distance with alignment constraints. In *Advances in Visual Computing*, p. 469–478. Springer.

VINTSYUK T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, **4**(1), 52–57.

WINKLER W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *ERIC*.