Analyse argumentative du corpus de l'ACL (ACL Anthology)

Untel Trucmuche[1, 2] Unetelle Machinchose[1, 3]
(1) LPL, AMU, CNRS, 5 avenue Pasteur, 13100 Aix-en-Provence
(2) LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9
(3) Lab, adresse, CP Ville, Pays
utrucmuche@lpl-aix.fr, umachinchose@adresse-academique.fr

Résumé. Cet article présente un essai d'application de l'analyse argumentative (*text zoning*) à l'ACL Anthology. Il s'agit ainsi de mieux caractériser le contenu des articles du domaine de la linguistique informatique afin de pouvoir en faire une analyse fine par la suite. Nous montrons que des technique récentes d'analyse argumentative fondées sur l'apprentissage faiblement supervisé permettent d'obtenir de bons résultats.

Abstract. This paper presents an application of Text Zoning to the ACL Anthology. Text Zoning is known to be useful to characterize the content of papers, especially in the scientific domain. We show that recent techniques based on weakly supervised learning obtain excellent results on the ACL Anthology. Although these kinds of techniques is known in the domain, it is the first time it is applied to the whole ACL Anthology.

Mots-clés : Analyse argumentative, corpus de textes scientifiques, ACL Anthology.

Keywords: Text Zoning, Corpus of scientific texts, ACL Anthology.

1 Introduction

L'analyse des masses de données (en anglais *big data*) est un thème de recherche porteur aujourd'hui. Les masses de données permettent en effet de mettre au jour des phénomènes difficilement observables sans méthodes automatiques, et la numérisation de tous les secteurs de la société permet aujourd'hui d'avoir accès à des données en grandes quantités pour un grand nombre de domaines.

La science est un des domaines qui produit ainsi de nombreuses données informatisées (littérature scientifique, mais aussi données brutes sous forme de textes, d'images, de chiffres, etc.) et la numérisation des données passées permet aujourd'hui d'avoir accès, pour plusieurs domaines très variées, à des collections d'articles de recherche s'étendant sur plusieurs dizaines d'années. Le monde de la linguistique informatique n'est pas en reste et l'ACL Anthology met aujourd'hui à la disposition des chercheurs plus de 24500 articles au format PDF. Les plus anciens articles datent de 1965 (première édition de COLING) mais ce n'est qu'à partir des années 1980 qu'on commence à avoir des données relativement conséquentes, le volume allant grandissant chaque année depuis lors (il y a donc une très grande disparité dans les volumes de données disponibles suivant les années considérées). Il existe des bases de données similaires pour la biologie et le domaine biomédical (par ex. Medline), les systèmes complexes ou la physique (par ex. APS data set de la *American Physical Society*) pour citer quelques bases ayant fait l'objet d'enquêtes diverses.

L'ACL Anthology a reçu un intérêt particulier en 2012 pour les 50 ans de l'*Association for Computational Linguistics*. Un atelier s'intitulant "*Rediscovering 50 Years of Discoveries*" a été organisé cette année-là (Banchs, 2012) : il s'agissait pour l'association de jeter un regard sur l'évolution du domaine depuis 50 ans. Au-delà de ces circonstances particulières, cet événement a été l'occasion d'analyser les données accumulées depuis 50 ans (mais pour les raisons données plus haut, la plupart des études portent sur les articles produits depuis 1980) avec les outils modernes issus à la fois du traitement des langues et des systèmes complexes, afin d'analyser l'évolution du domaine. Ce type d'analyse reste toutefois superficiel si une analyse fine de la structure des articles ou de leur résumé n'est pas effectuée au préalable.

La reconnaissance et l'annotation de la structure discursive des articles scientifiques présentent en effet des enjeux importants pour la communauté scientifique mais aussi pour le monde de l'édition. Ce type de techniques peut en effet permettre

de savoir si une section d'un article scientifique donné concerne, par exemple, le protocole expérimental employé, les données d'expériences ou la discussion et la comparaison avec les travaux antérieurs. Ce type d'analyse donne des résultats de plus en plus précis et commence à intéresser les grandes maisons d'éditions scientifiques, dans la mesure où on peut ainsi enrichir les bases de connaissances existantes et proposer de nouveaux parcours de lecture.

Nous présentons dans cet article une application de la reconnaissance automatique de la structure argumentative à l'ensemble des articles du corpus de l'ACL (*ACL Anthology*). Notre but à terme est d'analyser ce corpus suivant différents plans : évolution des thèmes de recherche au cours du temps, positionnement des acteurs dans le domaine, évolution des techniques utilisées, etc. Il est du coup primordial de pouvoir caractériser la structure des articles et/ou de leur résumé : les citations, les mentions des techniques employées ou des méthodes d'évaluation sont très dépendantes de leur contexte d'emploi comme l'ont démontré les travaux de S. Teufel depuis une quinzaine d'années. Il est par exemple essentiel de savoir si une technique d'analyse est citée dans les travaux antérieurs ou est réellement celle utilisée par l'auteur de l'article.

Nous présentons un rapide état de l'art présentant les évolutions récentes de l'annotation argumentative. Nous présentons ensuite l'application de ce type de techniques à l'ACL Anthology. La section suivante est consacrée aux résultats et à la discussion, avant la conclusion.

2 Etat de l'art

Les premiers travaux d'importance en matière d'analyse de la structure rhétorique sont certainement ceux de Simone Teufel (Teufel, 1999) qui a proposé de catégoriser les phrases d'articles de traitement automatique des langues suivant sept étiquettes différentes : BKG (arrière-plan scientifique), OTH (description neutre de travaux antérieurs), OWN (description neutre du travail de l'auteur), AIM (objectifs de l'article), TXT (annonce de l'organisation de l'article), CTR (comparaison avec des travaux antérieurs) et BAS (description des travaux antérieurs sur lesquels s'appuie l'article).

La tâche est appelée « rhetorical zoning » ou « argumentative zoning » par l'auteur, dans la mesure où le balisage doit permettre d'identifier la fonction rhétorique ou argumentative de chaque phrase du texte.

Le travail initial de S. Teufel (Teufel, 1999) est fondé sur l'annotation manuelle de 200 articles représentatifs du domaine issus des conférences de l'ACL et de la revue *Computational Linguistics*. Un classifieur est ensuite entraîné sur cette base : il permet d'obtenir une annotation automatique de nouveaux textes donnés en entrée à l'analyseur. L'auteur rapporte que le système automatique donne le bon résultat dans 70% des cas, comparé à un accord de 88% entre humains. Le classifieur repose sur un modèle bayesien naïf car les méthodes plus sophistiquées testées par l'auteur ne semblent pas permettre d'obtenir de meilleurs résultats.

Teufel montre dans une publication ultérieure (Teufel & Moens, 2002) comment cette technique peut être utilisée pour générer des résumés automatiques de qualité. Les techniques de résumé traditionnelles sont fondées sur la sélection de phrases en fonction de leur intérêt informatif supposé, essentiellement sur la base des noms et des verbes qui la compose (les mots les plus centraux, souvent appelés centroïdes (Radev *et al.*, 2004)), ce qui pose problème pour générer des textes tenant compte de la variété du texte de départ. Le repérage de la structure argumentative répond partiellement à ce problème dans la mesure où il est dès lors possible de générer des résumés reflétant les différentes zones repérées ou, au contraire, privilégiant une zone donnée suivant les besoins informationnels du lecteur.

Teufel a enfin montré (Teufel *et al.*, 2006) comment le marquage argumentatif peut être couplé avec les références scientifiques. Les articles scientifiques sont en effet fondées sur des citations des travaux antérieurs mais ces citations peuvent avoir différentes finalités : simple mention de travaux antérieurs donnant l'arrière-plan de la recherche en cours, travaux précis auxquels s'oppose la publication en cours, référence à des travaux utilisant le même protocole expérimental, etc. Coupler repérage de référence et balisage argumentatif permet de typer les citations, toujours dans le but de faciliter la lecture en fonction des besoins informationnels du lecteur.

Les travaux de S. Teufel ont depuis donné lieu à différents types de travaux, d'une part pour affiner la méthode d'annotation, d'autre part pour vérifier son applicabilité à différents domaines scientifiques. Pour le premier point, les recherches ont porté sur les traits pertinents pour la classification, l'évaluation de différents algorithmes pour la tâche et surtout la diminution de la quantité de texte à annoter pour obtenir un système fonctionnel. Pour le second, c'est surtout le domaine de la biomédecine et de la biologie qui ont montré le plus d'intérêt pour ce type de techniques, du fait de la quantité d'articles disponible dans ce domaine et de la nécessité d'accéder de manière transversale à cette littérature (les biologistes peuvent par exemple avoir besoin d'accéder à tous les protocoles expérimentaux pour un problème donné) (Mizuta *et al.*,

2006; Tbahriti *et al.*, 2006).

Les travaux de Y. Guo (Guo *et al.*, 2011, 2013) reprennent l'analyse de la structure argumentative en complétant les travaux initiaux de S. Teufel sur un certain nombre de points : recours à une vaste liste de critères pour déterminer la classification des phrases, évaluation de plusieurs algorithmes d'apprentissage et diminution de la quantité de données annotées à fournir au système pour l'entraînement.

Y. Guo *et al.* (2011) proposent en particulier d'avoir recours à l'apprentissage peu supervisé pour entraîner leur système. On sait en effet que ce type d'approche permet de réduire la quantité de données annotées en utilisant parallèlement une grande masse de données non annotées : cette méthode est bien indiqué dans notre cas dans la mesure où les corpus à analyser en traitement des langues (et particulièrement l'ACL Anthology) sont souvent d'assez grande taille mais qu'il ne sont évidemment pas annotés. Les traits utilisés pour l'apprentissage sont de trois types : *i*) positionnels (localisation de la phrase au sein de l'article), *ii*) lexicaux (mots, classes de mots, bigrammes, etc. sont pris en considération) et *iii*) syntaxiques (les différentes relations syntaxiques, ainsi que les classes de noms en position sujet et les classes de noms en position objet sont pris en considération). L'analyse est donc considérablement plus riche que celle de Teufel mais nécessite en contrepartie un analyseur syntaxique.

Au niveau du modèle d'apprentissage, Guo et al. (2011) propose une méthode originale fondée sur l'apprentissage actif. Pour l'apprentissage, les SVM actifs (*Active SVM*) se fondent sur un nombre limité d'exemples annotés mais sont capables de repérer des configurations où l'algorithme risque de donner des résultats incertains. Ces configurations particulières doivent alors faire l'objet d'une annotation manuelle, ce qui permet à l'algorithme de progresser rapidement en se focalisant uniquement sur les cas limites en matière de catégorisation. Le repérage des données à annoter est souvent fondé sur la structure de l'apprentissage au moyen de SVM : par exemple, les points qui sont à la frontière de l'hyperplan permettant la catégorisation sont souvent pris en considération en priorité (Tong & Koller, 2001; Novak *et al.*, 2006).

Dans le cadre de notre expérience, les configurations particulières faisant l'objet d'un étiquetage manuel sont ceux pour lesquels l'algorithme ne distingue pas de classe majoritaire de manière claire. Ce sont les phrases ayant reçu une probabilité proche entre les différentes classes possibles qui devront faire l'objet d'un étiquetage manuel.

3 Application de l'analyse argumentative au corpus de l'ACL

La méthode développée par Y. Guo et ses collègues semble particulièrement bien adaptée à notre problème. Nous souhaitons en effet catégoriser les termes repérés à l'étape précédente afin notamment d'identifier les méthodes mentionnées dans le corpus ACL Anthology et pouvoir ainsi analyser, par exemple, leur évolution dans le temps. Les termes apparaissant dans des phrases se rapportant au protocole expérimental employé sont donc susceptibles de particulièrement nous intéresser. Il faut noter à ce propos qu'il n'y a pas de frontière étanche entre thèmes et méthodes de recherche dans la mesure où le traitement automatique des langues s'appuie sur ses propres résultats pour concevoir des systèmes en couches empilées : ainsi, un analyseur sémantique reposera fréquemment sur un analyseur syntaxique employé comme outil (et apparaissant donc dans la section méthodologique de l'article).

L'annotation ne porte que sur les résumés des articles. On fait en effet l'hypothèse que les résumés contiennent assez d'information et sont assez redondants pour observer l'évolution du domaine. A l'inverse, aborder l'étude en utilisant le texte complet des articles entraînerait probablement du bruit et complexifierait inutilement les traitements.

Le jeu d'annotation initialement adopté comporte sept catégories différentes et une catégorie AUTRE pour les phrases ne pouvant pas être catégorisées par les étiquettes définies. Ces étiquettes sont les suivantes :

— OBJECTIF : décrit les objectifs de l'article ;
— METHODE : méthodes employées par l'article ;
— RESULTATS : résultats obtenus ;
— CONCLUSION : conclusion de l'article ;
— ARRIERE-PLAN : contexte scientifique ;
— TRAVAUX LIES : positionnement par rapport à des travaux directement liés à ceux présentés ;
— AUTRES TRAVAUX : positionnement par rapport à d'autres travaux.

Ces catégories sont reprises des travaux précédents, notamment (Mizuta *et al.*, 2006; Guo *et al.*, 2011, 2013). Il nous a semblé important de reprendre un jeu de catégories existantes dans la mesure où ces catégories, avec de légères variations, se sont globalement imposées depuis les premiers travaux de S. Teufel. Certaines catégories sont malgré tout peu présentes

dans les résumés de l'ACL Anthology, et finalement quatre catégories transparaissent principalement : les catégories OBJECTIF, ARRIERE-PLAN, RESULTATS et METHODE. Il est rare de trouver des comparaisons avec d'autres travaux dans les résumés de l'ACL Anthology (alors qu'on en trouve fréquemment dans les résumés en biologie par exemple).

Une centaine de résumés d'article issus de l'ACL Anthology ont ensuite été annotés manuellement avec ces catégories (environ 500 phrases, les résumés de l'ACL Anthology étant souvent très courts dans la mesure où il s'agit en grande majorité de résumés d'articles de conférence). Les articles annotés ont été choisis aléatoirement, en prenant soin toutefois qu'ils couvrent différentes périodes et qu'ils contiennent des termes variés. L'annotation a été faite en suivant le guide d'annotation mis au point par Y. Guo, notamment en ce qui concerne les phrases complexes, se rapportant potentiellement à plus d'une catégorie définie (un jeu de préférences est défini pour résoudre ces cas difficiles).

L'algorithme de (Guo *et al.*, 2011) est ensuite repris et adapté à notre cas de figure. L'analyse se fonde en particulier sur les traits positionnels, lexicaux et syntaxiques comme expliqué dans la section précédente. Aucune information spécifique au domaine n'est ajoutée, ce qui rend le processus simple à modéliser et à reproduire. Pour l'analyse syntaxique, le Stanford Parser est utilisé : `http://nlp.stanford.edu/software/lex-parser.shtml` (De Marneffe *et al.*, 2006). On utilise le SVM linéaire tel qu'implémentée dans Weka pour la classification, avec le paramètre -M pour le calcul des probabilités post-apprentissage. On pourra se reporter à (Guo *et al.*, 2011) pour les détails de l'algorithme d'apprentissage et de marquage des phrase selon le modèle défini. Comme résultat, pour chaque phrase du corpus, l'algorithme associe une étiquette choisie parmi les étiquettes possibles.

4 Résultats et discussion

Pour valider les résultats obtenus, un ensemble de résumés est choisi aléatoirement. Les quatre catégories principales sont bien représentées mais inégalement réparties : 18,05 % des phrases sont catégorisées comme ARRIERE-PLAN, 14,35 % comme OBJECTIF, 14,81 % comme RESULTAT et 52,77 %comme METHODE. On voit bien, à la lecture de ces chiffres, l'importance de la dimension méthodologique dans le domaine.

On observe ensuite, pour chaque catégorie possible, le pourcentage de phrases correspondant effectivement à cette étiquette, ce qui permet de mesurer les performances du système en terme de précision. Les résultats obtenus sont présentés dans le tableau suivant.

TABLE 1 – Résultat de l'analyse argumentative (en précision)

Catégorie	Précision
Objectif	83,87 %
Arrière-plan	81,25 %
Méthode	71,05 %
Résultats	82,05 %

Ces résultats sont conformes à l'état de l'art. On voit que les résultats sont globalement satisfaisants, particulièrement en regard du peu de phrases annotées pour l'entraînement. La richesse des traits pris en compte et la stratégie d'apprentissage actif permettent en outre d'avoir des résultats portables d'un domaine à l'autre sans tâche d'annotation lourde. Les résultats sont légèrement moins bons pour la catégorie METHODE car celle-ci est sans doute plus diversifiée que les autres et donc moins facile à cerner.

L'exemple 1 est un texte annoté suite à l'analyse du système (il s'agit de l'article de (Lee *et al.*, 2002), choisi au hasard parmi ceux qui présentent une bonne diversité dans les catégories utilisées). La catégorisation s'effectue au niveau des phrases, ce qui n'est pas sans poser problème : par exemple, dans ce résumé, le fait qu'une méthode hybride est utilisée est indiqué dans une phrase étiquetée OBJECTIF par le système. Les phrases marquées METHODE contiennent toutefois des mots clés précieux, comme *lexical pattern* ou *tri-gram estimation*, ce qui peut permettre d'inférer le fait qu'il s'agit d'un système hybride. On aperçoit au passage des problèmes de numérisation, qui sont typiques du corpus étudié : l'ACL Anthology comprend des textes convertis automatiquement à partir de fichiers PDF de conférences passées, ce qui entraîne parfois des problèmes de qualité.

Exemple 1 : *Un résumé annoté avec l'analyseur de la structure argumentative. Les catégories ajoutées au texte sont*

indiquées en gras.

> Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes . **ARRIERE-PLAN**
> This paper presents a generalized unknown morpheme handling method with POSTAG(POStech TAGger) which is a statistical/rule based hybrid POS tagging system . **OBJECTIF**
> The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation . **METHODE**
> The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result . **METHODE**
> In our scheme , we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol , which was not possible before in Korean tagging systems . **RESULTATS**
> In a series of experiments using three different domain corpora , we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora . **RESULTATS**

Il est dès lors possible d'extraire des mots clés puis de classer ces mots clés suivant les zones considérées. Pour donner un exemple simple d'analyse reposant sur la catégorisation effectuée ici, on peut essayer de tracer l'évolution de différentes tendances dans le temps. Par exemple, pendant la période considérée, les méthodes utilisées ont beaucoup changé, le principal fait marquant étant peut-être le recours massif à l'apprentissage depuis la fin des années 1990. Cette tendance est marquée par un recours quasi systématique dans les articles actuels à des expérimentations donnant lieu à des résultats chiffrés.

Pour confirmer de façon quantitative cette hypothèse, nous nous intéressons à l'évolution dans le temps de la proportion de phrases étiquetées RESULTAT. Sur la figure 1, nous pouvons ainsi observer que la courbe correspondante croit de façon quasi linéaire du début des années 1980 jusqu'à la fin des années 2000.

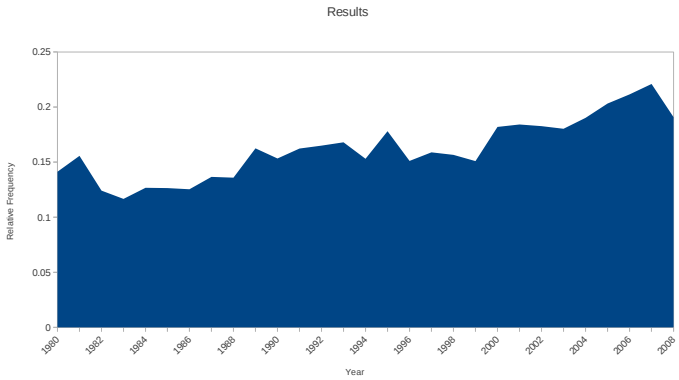

FIGURE 1 – Évolution dans le temps de la proportion de phrases catégorisées par l'outil d'analyse discursive comme étant des phrases concernant des résultats (par rapport au nombre total des phrases contenues dans les articles publiés dans l'année correspondante

Cette illustration relativement simple n'est donnée ici qu'à titre d'exemple. Plus globalement, l'analyse contenue dans cet article ouvre par exemple des perspectives pour une analyse fine de l'évolution des méthodes réellement employées dans les articles, en se focalisant sur la zone étiquetée METHODE (tout en excluant les section ARRIERE-PLAN). Une analyse plus fouillée de l'ACL Anthology en prenant en compte l'analyse argumentative reste à faire, à la suite des travaux de (Anderson *et al.*, 2012) pour les 50 ans de l'ACL.

5 conclusion

Cet article a permis d'appliquer l'analyse argumentative (*text zoning*) à l'ensemble de l'ACL Anthology avec des résultats satisfaisants bien qu'une partie tout à fait minime du corpus initial ait fait l'objet d'une annotation manuelle. On voit ainsi que les techniques d'analyse argumentative, surtout appliquées au domaine bio-médical ces dernières années, sont aussi utilisables dans d'autres cadres et pour d'autres domaines (conformément aux travaux pionniers de S. Teufel) et restent efficaces même pour de grandes masses de documents. Ce type d'analyse devrait permettre d'affiner l'étude des grands corpus scientifiques tant sur le plan historique qu'épistémologique.

Références

ANDERSON A., JURAFSKY D. & MCFARLAND D. A. (2012). Towards a computational history of the acl : 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, p. 13–21, Jeju Island, Corée : Association for Computational Linguistics.

R. E. BANCHS, Ed. (2012). *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Jeju Island, Corée : Association for Computational Linguistics.

DE MARNEFFE M.-C., MACCARTNEY B. & MANNING C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the conference on Language Resource and Evaluation (LREC)*, p. 449–454.

GUO Y., KORHONEN A. & POIBEAU T. (2011). A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 273–283, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

GUO Y., REICHART R. & KORHONEN A. (2013). Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, p. 928–937.

LEE G. G., LEE J.-H. & CHA J. (2002). Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of korean. *Computational Linguistics*, **28**(1), 53–70.

MIZUTA Y., KORHONEN A., MULLEN T. & COLLIER N. (2006). Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, **75**(6), 468–487.

NOVAK B., MLADENI D. & GROBELNIK M. (2006). Text classi ?cation with active learning. In *From Data and Information Analysis to Knowledge Engineering*, p. 398–405.

RADEV D. R., JING H., STYŚ M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, **40**(6), 919–938.

TBAHRITI I., CHICHESTER C., LISACEK F. & RUCH P. (2006). Using argumentation to retrieve articles with similar citations : An inquiry into improving related articles search in the medline digital library. *I. J. Medical Informatics*, **75**(6), 488–495.

TEUFEL S. (1999). *Argumentative Zoning : Information Extraction from Scientific Articles*. University of Edinburgh.

TEUFEL S. & MOENS M. (2002). Summarizing scientific articles – experiments with relevance and rhetorical status. *Computational Linguistics*, **4**(28), 409–445.

TEUFEL S., SIDDHARTHAN A. & TIDHAR D. (2006). Automatic classification of citation function. In *Proceedings of Empirical Methods in Natural language Processing (EMNLP)*, p. 103–110 : Association for Computational Linguistics.

TONG S. & KOLLER D. (2001). Active Learning with Applications to Text Classi ?cation. *Journal of Machine Learning Research*, **2**, 45–66.