MULTITAG, une ressource linguistique produit du paradigme d'évaluation

Patrick Paroubek & Martin Rajman

LIMSI - CNRS, Batiment 508 Universite Paris XI, 91403 Orsay Cedex
Email: pap@limsi.fr, URL:http://www.limsi.fr/Individus/pap
&
Laboratoire d'Intelligence Artificielle, Département Informatique École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne-Ecublens, Switzerland
Email:rajman@lia.di.epfl.ch, URL:http://liawww.epfl.ch/People/rajman.html

Texte

Résumé

Dans cet article, nous montrons comment le paradigme d'évaluation peut servir pour produire de façon plus économique des ressources linguistiques validées de grande qualité. Tous d'abord nous présentons le paradigme d'évaluation et rappelons les points essentiels de son histoire pour le traitement automatique des langues, depuis les premières applications dans le cadre des campagnes d'évaluation américaines organisées par le NIST et le DARPA jusqu'aux derniers efforts européens en la matière (SENSEVAL2/ROMANSEVAL2, CLEF, CLASS etc.). Nous présentons ensuite le principe qui permet de produire à coût réduit des ressources linguistiques validées et de grande qualité à partir des données qui sont produites lorsque l'on applique le paradigme d'évaluation. Ce principe trouve ses origines dans les expériences (Recognizer Output Voting Error Recognition) qui ont été effectuées pendant les campagnes d'évaluation américaine pour la reconnaissance automatique de la parole. Il consiste à combiner les données produites par les systèmes à l'aide d'une simple stratégie de vote pour diminuer le nombre d'erreurs. Nous faisons alors un lien avec les stratégies d'apprentissages automatiques fondées sur la combinaison de systèmes de même nature. Notre propos est illustré par la description de la production du corpus MULTITAG (projet du programme Ingénierie des Langues des département SPI et SHS du CNRS) à partir des données qui avaient été annotées lors de la campagne d'évaluation GRACE, correspondant à un corpus d'environ 1 million de mots annotés avec un jeu d'étiquettes morpho-syntaxiques de grain très fin dérivé de celui qui a été défini dans les projets EAGLES et MULTEXT. Nous présentons le corpus MULTITAG et la procédure qui a été suivie pour sa production et sa validation. Nous concluons en présentant le gain obtenu par rapport à une méthode classique de validation de marquage morho-syntaxique.

1. Le paradigme d'évaluation

Bref historique. L'évaluation comparative a été utilisée depuis 1984 dans le programme Américain du DARPA comme un paradigme de base pour l'ingénierie linguistique. En de-

hors des États-Unis, des activité similaires se sont développées, tant au niveau national qu'au niveau européen, mais à une échelle plus réduite et sur des périodes de temps plus courtes (pas de programme d'envergure s'étendant sur plusieurs années voir une décennie) (Mariani & Paroubek, 1999). Les efforts européens récents en matière d'évaluation pour les technologies du langage sont représentés par les projets CLEF (évaluation de systèmes de recherche documentaire interlingue, en collaboration avec le NIST les conférences TREC), la prochaine campagne d'évaluation SENSEVAL-2/ROMANSEVAL-2 (Kilgarriff, 1998), le projet européen CLASS (action transversale de sensibilisation aux problèmes d'évaluation au sein de groupes de projets du cinquième programme cadre de la Communauté Européenne). Au plan national, on notera la seconde campagne d'évaluation Amaryllis/ARC-A1 (Coret *et al.*, 1997)[1] (financée par le Ministère de la Recherche et l'AUF dans le cadre des ARC du réseau FRANCIL) qui traite de la recherche documentaire en langue française, avec un protocole inspiré des campagnes TREC, mais avec des spécificités propres, en particulier en ce qui concerne la taille des documents et la méthode de construction du référentiel. Cette campagne est récemment arrivée à son terme et, suite au succès rencontré, sa prochaine édition a déjà été annoncée.

Principes généraux. L'évaluation comparative consiste en la comparaison par une ensemble de participants et sur des données communes des systèmes qu'ils ont mis en oeuvre sur une même tâche de contrôle définie à l'avance (Bernsen *et al.*, 1999). Les performances des systèmes pour la réalisation de la tâche sont mesurées par une métrique et un protocole validés à l'avance par les participants. Ainsi définie, l'évaluation comparative implique : (1) la définition de la tâche de contrôle, (2) l'identification des développeurs ou des intégrateurs intéressés à tester leur système sur la tâche sélectionnée, (3) l'organisation de la campagne d'évaluation avec, en particulier, la distribution de données linguistiques pour l'entraînement et le test des systèmes, et finalement (4) la définition des mesures et du protocole qui seront utilisés pour évaluer les performances des systèmes. Notons que la tâche de contrôle définit non seulement le traitement que les systèmes doivent réaliser pendant l'évaluation, mais aussi les conditions dans lesquelles ce traitement doit s'effectuer. Par exemple, pour évaluer des analyseurs syntaxiques, on peut demander aux analyseurs de produire des analyses complètes ou simplement de se contenter d'identifier les frontières de syntagmes. Cependant, quelles que soient les spécificités de mise en oeuvre, l'application du paradigme d'évaluation implique la production de données langagières. De plus, lorsque le paradigme d'évaluation utilise une méthodologie d'évaluation quantitative boîte noire (Sparck-Jones & Galliers, 1995), les données produites sont en général abondantes : les organisateurs construisent des données de test et de référence et les participants utilisent leurs systèmes pour produire les données qui serviront mesurer les performances des systèmes. Ces données pourraient donc facilement être réutilisées comme matériaux d'apprentissage si les coût de correction des erreurs faites par les systèmes et ceux la validation finale des données n'étaient pas si élevés.

2. Combiner pour améliorer

En apprentissage automatique, il est désormais bien établi que la combinaison d'un ensemble de méthode ou de systèmes d'apprentissage permet souvent d'obtenir de meilleurs résultats qu'avec une méthode ou un système pris isolément. Par exemple, un des algorithmes très prometteurs qui fait actuellement l'objet de travaux de recherche importants est l'algorithme *Ada boost* (Schwenk, 1999) qui consiste à entraîner une cascade de systèmes similaires, chacun étant

1. voir l'URL: http://www.inist.fr/accueil/profran.htm

chargé de traiter les erreurs laissées par les systèmes précédents. Si l'on remonte plus loin dans le temps, on trouve de nombreuses utilisations de la stratégie du "gagnant ramasse tout" (*winner take all*) qui consiste à mettre en compétition, au sein d'un système, la sortie de plusieurs unités de traitement similaires (Simpson, 1990).

Par ailleurs, au cours du programme d'évaluation sur la reconnaissance de la parole (S., 1998), le NIST a développé la méthode ROVER (*Recognizer Output Voting Error Reduction*) (Fiscus, 1997) qui permet de produire, par combinaison automatique, une transcription d'un signal de parole à partir des transcriptions fournies par les différents systèmes de reconnaissance participant à l'évaluation. Le système composite résultant a toujours eu de meilleurs performances (en précision) que n'importe lequel de ses constituants élémentaires. Dans l'approche ROVER, la sortie des systèmes de reconnaissance de la parole est d'abord combinée en un unique graphe de mots au moyen d'une version modifiée de l'algorithme d'alignement (programmation dynamique) utilisé par le NIST pour mesurer les performances des systèmes[2]. Ce graphe est ensuite élagué à l'aide d'une simple stratégie de vote (à la majorité simple) qui permet de sélectionner le meilleur choix à chaque point de décision. Plus précisément, les votes s'appuient sur les fréquences maximales d'occurrence et une valeur seuil du score de confiance (les sorties des systèmes de transcription avaient été annotées avec un score de confiance (Chase, 1997)). Dans ces conditions, (Fiscus, 1997) indique une réduction de l'erreur de transcription des mots de 5.6% en absolu et de 12.5% en relatif.

Pour ce qui est de l'annotation morpho-syntaxique, le principe de la combinaison de systèmes a été utilisé dans le passé par (Marquez & Padro, 1998) qui ont combiné deux systèmes d'annotation pour marquer un corpus et par (Tufis, 1999) qui a proposé d'utiliser plusieurs versions d'un même système, chacune entraînée sur des données de type différent.

3. Les données produites lors de la campagne d'évaluation GRACE

GRACE (Adda *et al*., 1999) a été la première campagne d'évaluation à grande échelle pour le marquage morpho-syntaxique du français écrit. C'était une initiative du programme national CCIIL (Cognition, Communication Intelligente et Ingénierie du Langage) des départements SHS et SPI du CNRS.

L'appel à participation a été publié en Novembre 1995 et la première année a servi à lancer l'action et à mettre en place les différents comités d'organisation. Du point de vue des participants, la campagne GRACE a été organisée en 3 phases : entraînement, essais et tests. La première a permis aux participants de calibrer leur système sur des données brutes non annotées, tandis que les deux phases suivantes ont consisté en une application complète du protocole d'évaluation en vraie grandeur sur des données réelles. Lors de ces phases, les participants devaient annoter d'une grande quantité de texte (plusieurs centaines de milliers de mots) et fournir une table de correspondance entre leur jeu d'étiquettes et le jeu d'étiquettes de référence qui avait été dérivé, en collaboration avec les participants, des propositions de standard publiées par EAGLES (Leech & Wilson, 1995) et raffinées dans le projet MULTEXT (Ide & Véronis, 1994).

Le corpus d'entraînement a été distribué simultanément à tous les participants en janvier 1996. Le corpus d'essai a été distribué individuellement à chaque participant (sous une forme encryptée) durant l'automne 1996. Les résultats des essais ont été présentés et discutés lors

2. La boîte à outils SCLITE utilisée à cet effet est disponible gratuitement à l'URL http://www.itl.nist.gov/iaui/894.01/software.htm

d'un atelier satellite lors des Journées Scientifiques et Techniques du Réseau FRANCIL en avril 1997 (Adda. *et al*., 1997). Les données de test ont été distribuées à la fin du mois de décembre 1997, de la même manière que les données d'essai. Les résultats préliminaires des tests ont été discutés avec les participants pendant une journée atelier en mai 1998 et les résultats finaux ont été rendus publics sur Internet durant l'automne 1998 [3], après avoir été validés par les participants et les organisateurs (procédure croisée utilisant 2 chaînes distinctes de validation implémentant deux algorithmes différents et réalisées par deux sites différents). Au début de l'action, il y avait 21 participants de 5 pays différents (CA, USA, D, CH, FR), provenant à la fois du secteur public et de l'industrie, et 3 évaluateurs : l'EPFL, l'INaLF-CNRS et le limsi-CNRS. Les deux fournisseurs de corpus étaient le limsi-CNRS et l'INaLF-CNRS. Parmi les 21 participants initiaux, 17 ont participé aux essais et 13 aux tests finaux. La taille du corpus d'entraînement était d'environ 10 millions de mots non annotés, réparti de façon équilibrée entre des textes littéraires et des textes journalistiques.

Pour les essais, les participants devaient annoter un corpus d'environ 450 000 mots avec une distribution en genre similaire à celle du corpus d'entraînement. Les mesures de performance ont été effectuées sur 20 000 mots pour lesquels une description morpho-syntaxique de référence avait été produite manuellement.

Pour les tests, les participants avaient à marquer un corpus d'environ 650 000 mots et les mesures ont été effectuées sur environ 40 000 mots. La métrique (mesure quantitative boîte noire) utilisée dans GRACE repose sur les notions de *Décision* et de *Précision* dérivées des mesures de Précision et de Rappel utilisées en recherche documentaire. Dans GRACE, la précision mesure la capacité d'un système à affecter en contexte une étiquette correcte à une forme donnée et la décision mesure la capacité d'un système à restreindre en contexte le nombre d'étiquettes possibles pour une forme donnée dans le jeu d'étiquettes utilisé. Une des leçons de l'expérience GRACE a été la mise en évidence de l'importance de la double production des résultats par deux chaînes d'évaluation distinctes. La boîte à outils d'évaluation utilisée a été raffiné dans le cadre du projet européen ELSE et est distribuée gratuitement [4].

La campagne GRACE s'est révélée être un succès et ses retombées ont été nombreuses : une meilleure connaissance des systèmes existants et de leur niveau de développement; une mesure d'évaluation précise définie en collaboration avec les participants; une boîte à outils d'évaluation disponible gratuitement; un nouveau produit sur le marché (suite aux résultats obtenus lors des essais, un participant industriel a décidé d'ajouter son système d'étiquetage à son catalogue); la création d'une communauté d'acteurs intéressés par l'évaluation; et enfin les données de base nécessaires pour construire la ressource linguistique présentée dans cet article.

4. MULTITAG

Objectifs du projet. Le projet MULTITAG (du programme de recherche Ingénierie des Langues, commun aux départements SHS et SPI) [5] avait pour objectif la production et la mise à disposition (distribution par ELRA) d'un corpus de 1 million de mots annotés par les informations morpho-syntaxiques extraites des corpus marqués par les participants lors de la campagne d'évaluation GRACE. Un tel corpus et sa documentation représentent non seulement un matériau intéressant pour les études linguistiques et une ressource importante pour l'entraînement

3. http://www.limsi.fr/TLP/grace

4. http://www.limsi.fr/TLP/ELSE

5. Les équipes impliquées dans MULTITAG ont été: l'INaLF-CNRS, le limsi-CNRS, le CILSH (U. Provence), et TALANA (U. Paris7)

des systèmes d'annotation morpho-syntaxique mais également un matériau de base très utile pour l'apprentissage automatique et l'évaluation de combinaisons de tels systèmes. Une première étude de ce corpus a déjà été réalisée dans (Illouz, 1999) où l'auteur présente des résultats préliminaires sur les différents types de texte (genre) qui composent le corpus et les variations des performances des systèmes d'annotation morpho-syntaxique associées. Une amélioration de la qualité de ce corpus a également été entreprise mais, pour réduire les coûts de correction et de validation, cette tâche a été réalisée de façon semi-automatique et seules les formes pour lesquelles les annotations des différents systèmes ne convergeaient pas (pas de vote majoritaire) ont été validées manuellement. Le degré d'accord entre les annotations produites par les différents systèmes a donc été utilisé comme une mesure de confiance permettant d'identifier les formes dont l'annotation avait prioritairement besoin d'une relecture.

Un exemple est donné dans la table 1 où, pour chaque étiquette, le vote (i.e. le nombre systèmes proposant cette étiquette) est indiqué entre accolades.

occ	tags & vote #
la	Dafsd{13}|Dafsi{4}|Damsd{1} |Damsi{1}|Ddfs{4}|Ddms{1} |Difs{4}|Dims{1}|Dkfs{1} |Drfs{1}|Ds1fsp{4}|Ds1fss{4} |Ds1msp{1}|Ds1mss{1}|Ds2fsp{4} |Ds2fss{4}|Ds2msp{1}|Ds2mss{1} |Ds3fsp{4}|Ds3fss{4}|Ds3msp{1} |Ds3mss{1}|Dtfs{2}|Dtms{1} |NULL{1}|Ncmp{1}|Ncms{2} |Pp3fsa{2}
couverture	NULL{1}|Ncfs{14}|Ncms{1}
du	Dai+Damsd{4}|Damsd{2} |NULL{1}|Sd{1}|Sp{2} |Sp+Damsd{10}

TAB. 1 – *Exemple de mesure de confiance pour les annotation morpho-syntaxique obtenue avec 15 systèmes*.

Le manque de disponibilité d'outils souples (i.e. facilement modifiables) et interactifs pour la correction de corpus annotés a incontestablement constitué un handicap pour le projet et il a finalement été décidé d'utiliser un tableur standard pour les relectures manuelles qui ont été faites en collaboration avec le CILSH (U. Provence) et TALANA (U. Paris7). Pour ce qui est des aspects plus linguistiques, la définition des instructions d'annotation pour les relecteurs a été beaucoup plus difficile qu'initialement prévu. De ce fait le projet MULTITAG a donc également permis, en plus de la validation des annotations du corpus et de définition de méthodes de combinaison de systèmes, de compléter le manuel d'annotation produit dans GRACE. Ce dernier sera très utile, non seulement pour d'autres opérations de validation (prise de décision et vérification de cohérence), mais aussi pour des études linguistiques plus générales. Il fait partie de la documentation du corpus.

Combinaison de systèmes d'étiquetage. Les techniques développées dans le cadre du projet MULTITAG ont été appliquées à la fois au corpus résultant de la phase d'essai et de la phase de test.

Le corpus d'essai a été normalisé[6]. Le résultat de la combinaison des systèmes a été produit mais aucune validation manuelle des résultats n'a été effectuée à ce jour. En effet, le corpus de test étant non seulement plus grand mais également plus pertinent car annoté avec la dernière version du jeu d'étiquettes GRACE, il a été décidé de donner la priorité à son traitement. Le corpus de test a donc tout d'abord été normalisé puis des mesures de performance ont été réalisées à l'aide des outils GRACE pour déterminer la meilleure combinaison de systèmes pour l'annotation. Les résultats obtenus ont montré qu'en comparaison avec les meilleurs résultats (en précision et en décision) obtenus individuellement par les participants, il était possible de produire une combinaison de systèmes ayant des performances meilleures sur l'une des dimensions (précision ou décision) sans dégradation notable sur l'autre (voir la figure 1). Ce résultat a été obtenu en combinant les annotations des 5 systèmes ayant les meilleurs résultats en précision (parmi les 15 systèmes, 13 participants et 2 systèmes utilisant une approche basique d'annotation). Il est à noter que les résultats obtenus en combinant l'ensemble des 15 systèmes ne montrent pas une telle amélioration globale (voir la table 2 et la figure 2), ce qui fait ressortir l'importance du choix des systèmes à combiner sur la base d'un critère (ici la précision) pertinent par rapport à la tâche de contrôle. D'après les rapports d'autres expériences de combinaison de système par vote, il semblerait que la meilleure stratégie consiste à prendre des systèmes utilisant des algorithmes de résolution différent mais ayant des performances voisines (pour le moment il s'agit encore d'un hypothèse).

Système	Précision	Décision	Précision Moyenne (à Déc. = 100)
Meilleure Préc.	97.9	25.6	40.4
Meilleure Déc.	94.8	100.0	94.8
Comb. 5 meill. Préc.	96.6	92.8	92.5
Comb. tous (15)	93.6	96.1	91.7

TAB. 2 – *Mesures de performances (en %) des deux meilleurs systèmes dans les 2 dimensions et de la combinaison respective de tous les sytèmes (15) et des 5 meilleurs systèmes en précision (voir figure 1).*

Traitement semi-automatique des ambiguïtés résiduelles. Pour ce qui est des formes pour lesquelles la procédure de combinaison ne produit pas de résultat univoque (vote équilibré entre plusieurs étiquettes), une désambiguïsation manuelle a été réalisée. Cette désambiguïsation a été effectuée en plusieurs étapes.

Tout d'abord, les 38 643 formes du corpus de test (soit environ 4% seulement de l'ensemble des 836 500 formes du corpus) pour lesquelles la procédure de combinaison avait produit un étiquetage ambigu pour la catégorie ou la sous-catégorie (indépendemment des autres informations morpho-syntaxiques comme le genre et le nombre) ont été traitées.

Puis, dans une seconde phase, l'information de genre, de nombre ou de personne a été validée manuellement pour toutes les formes concernées, soit 64 061 formes, ce qui représente environ 8% du corpus (voir la table 4 pour un exemple de résultat d'annotation à l'issue des phases 1 et 2).

Enfin, dans une troisième et dernière phase, les descriptions morpho-syntaxiques complètes

6. La normalisation correspond ici à la projection, à l'aide des tables de correspondances fournies et d'un algorithme de segmentation ("tokenisation") uniforme, des jeux d'étiquettes utilisés par les systèmes vers le jeu d'étiquettes de référence GRACE.

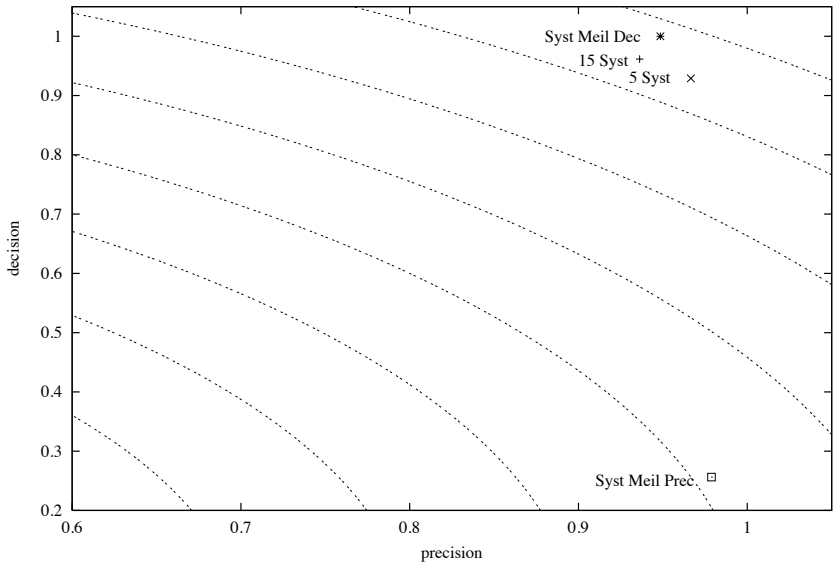

FIG. 1 – *Points dans le plan Précision/Décision représentant les mesures de performances des deux meilleurs systèmes dans les 2 dimensions et de la combinaison respective de tous les sytèmes (15) et des 5 meilleurs systèmes en précision (voir table 2).*

occ.	tags	check ?	Phase 1	Phase 2
Né	Afpms|Vmps-sm	1	Vm	ms-
à	Sp	0	-	- - -
Tarbes	Npfs|Npms	1	-	[mf]s-
,	F	0	-	- - -

TAB. 3 – *Exemples de correction manuelles des annotations morpho-syntaxiques en phase 1 et 2 du projet MULTITAG.*

ont été produites automatiquement en tenant compte des informations de catégorie, de sous-catégorie, de genre, de nombre et de personne produites manuellement. Au total, seules 85 442 (soit environ 10%) des formes initialement annotées par combinaison ont dû être modifiés par le biais d'opérations de désambiguïsation semi-automatiques. La première version de corpus ainsi produite a été transmise a ELRA pour distribution.

Validation. À titre indicatif, une première estimation effectuée sur un échantillon de 511 formes sélectionnées aléatoirement dans le corpus, situe la probabilité d'erreur résiduelle sur l'étiquetage des formes à 7.24% +/- 2.25% avec un coefficient de confiance à 95car 37 erreurs ont été relevées. Ce taux d'erreur est tout à fait satisfaisant par rapport au taux observé avec une annotations purement manuelles, compte tenu du fait que le travail correspondant est environ 10 fois moins important (relecture de seulement 10En ce qui concerne les performance pure, le taux d'erreur résiduel est aussi situé dans une plage de performance supérieure ou égale aux meilleures performances des systèmes telle que mesurée dans GRACE (avec le jeux d'etiquettes GRACE complet, i.e. une décision à 1, la meilleure performance en précision mesurée lors de l'évaluaiton GRACE a été de 94.8).

Par ailleurs, les mesures ont permis de valider l'un des intérêts supplémentaire de l'approche

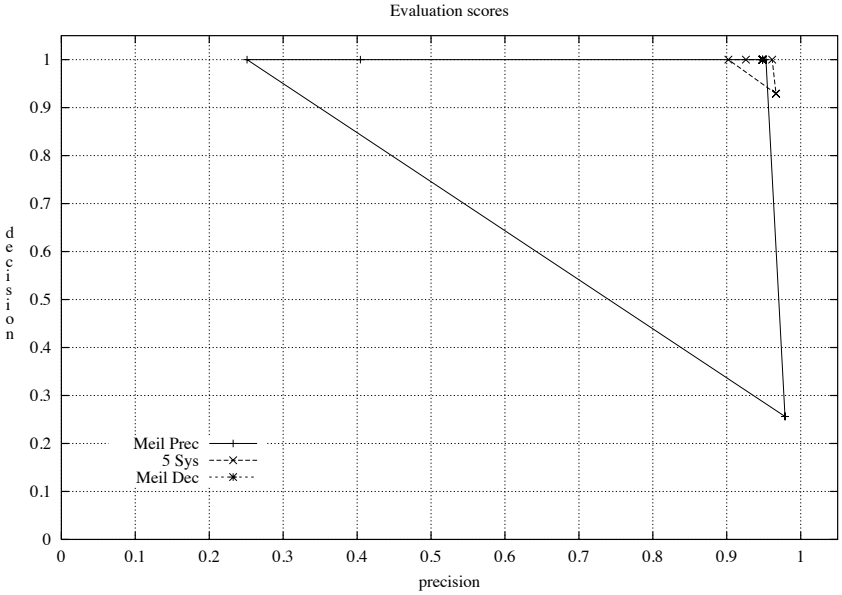

FIG. 2 – *Triangles représentant les zones de performance en Precision/Decision respectivement : du système ayant obtenu la meilleur valeur en Précision (grand triangle), du système ayant la meilleur valeur en Décision (petite étoile dans le coin en haut à droite) et du système résultant de la combinaison des 5 systèmes ayant eu les meilleurs résultats en Précision (petit triangle en haut à droite).*

à base de combinaison de système, à savoir l'identification des formes dont l'etiquettage est potentiellement erroné, en fonction du degré d'accord entre les systèmes. Le nombre de votes reçus par une étiquette fonctionne comme un indicateur de confiance. Parmi les 511 formes sélectionnées aléatoirement, il y avait 53 formes identifiées par la procédure de combinaison comme ayant une annotation incertaine et nous voyons que 27 d'entre-elles n'étaient pas correctement étiquetées (soit par ce que le relecteur avait fait une erreur, soit parce que les traits résiduels, c.a.d. autres que catégorie, sous-catégorie, genre, nombre et personne étaient ambigüs ou erronés), ce qui correspond à un taux d'erreur de 50.9% +/-13.5% avec un coefficient de confiance à 95d'après la procédure de combinaison, n'avaient pas besoin d'être revues, seulement 10 avaient en fait un étiquettage erroné, ce qui représente un taux d'erreur résiduel de 2.18% +/-1.34% avec un coefficient de confiance à 95à tous les taux d'erreur mesurés dans l'evaluation GRACE.

5. Conclusion

La campagne GRACE et l'expérience MULTITAG ont prouvé que le paradigme d'évaluation, lorsqu'il utilise des méthodes boîte noire quantitatives peut aussi servir à produire de manière économique des ressources linguistiques validées de grande qualité. Une telle approche peut facilement se généraliser à d'autres tâches de contrôle et fournir ainsi un moyen de rentabiliser les coûts du déploiement d'une campagne d'évaluation par la valorisation des produits dérivés, éliminant de la même façon, au moins partiellement, un des arguments usuellement présentés par les détracteurs de ce genre d'approche. Pour le corpus MULTITAG, les travaux futurs vont concerner la production d'annotations complémentaires (lemmes) selon la même procédure (des mesures préliminaires indiquent que seulement 10% des annotations de lemmes

occ. #	occ.	tag	Correction manuelle?
00007	mes	Ds1fps	0
00008	fonctions	Ncfp	0
00009	m	Pp1msa/1.2	1
00010	’	Pp1msa/2.2	1
00011	ont	Vaip3p	0
00012	successivement	Rgp	0
00013	appelé	Vmpssm	0
00014	à	Sp	0

TAB. 4 – *Un exemple des données produite par la phase finale de validation.*

fournies par les participants diffèrent), l'affinage de l'estimation des erreurs résiduelles, la vérification automatique de la cohérence du marquage et la levée semi-automatique des ambiguïtés résiduelles (traits morpho-syntaxique fins).

Références

ADDA. G., LECOMTE J., MARIANI J., PAROUBEK P. & RAJMAN M. (1997). Les procédures de mesure automatique de laction grace pour lévaluation des assignateurs de parties du discours pour le français. In *1 ères Journées Scientifiques et Techniques du Réseau Francophone de lIngenierie de la Langue de lAupelf-Uref*, Avignon.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M. & LECOMTE J. (1999). L'action grace d'évaluation de l'assignation des parties du discours pour le français. *Langues*, **2**(2), 119–129.

BERNSEN N.-O., CALZOLARI N., CHANOD J.-P., CHOUKRI K., DYBKJÆR L., GAIZAUSKAS R., KRAUWER S., LAMBERTERIE I., MARIANI J., NETTER K., PAROUBEK P., POPESCU-BELIS A., RAJMAN M. & ZAMPOLLI A. (1999). *A Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment*. Deliverable 1.1, EU project LE4-8340, Evaluation in Language and Speech Engineering. http://www.limsi.fr/TLP/ELSE/ELSED11EN.HTML.

CHASE L. (1997). Word and acoustic confidence annotation for large vocabulary speech recognition. In *European Conference On Speech Communication And Technology (Eurospeech)*, p. 815–818, Rohdes, Greece.

CORET A., KREMER P., LANDI B., SCHIBLER D., SCHMIDT L. & VISCOGLIOSI N. (1997). Accès à l'information textuelle en français: Le cycle exploratoire amaryllis. In *1 ères Journées Scientifiques et Techniques du Réseau Francophone de lIngenierie de la Langue de lAupelf-Uref*, Avignon.

FISCUS J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *IEEE Workshop on Automatic Speech Recongition and Understanding*.

IDE N. & VÉRONIS J. (1994). Multext: Multilingual text tools and corpora. In *15th International conference on computational linguistics (COLING)*, p. 588–592, Kyoto, Japan.

ILLOUZ G. (1999). Méta-Étiqueteur adaptatif, vers une utilisation pragmatique des ressources linguistiques. In *Conférence Traitement Automatique du Langage Naturel*, p. 185–194, Cargèse, France.

KILGARRIFF A. (1998). Senseval: An exercise in evaluating word sense disambiguation programs. In *1st International Conference on Language Resources and Evaluation (LREC98)*, volume 1, p. 581–585, Granada, Spain.

LEECH G. & WILSON A. (1995). *EAGLES morphosyntactic annotation - EAG-CSG/IR-T3.1*. Rapport interne, Consiglio Nazionale delle Ricerche, Instituto di Linguistica Computazionale, Pisa.

MARIANI J. & PAROUBEK P. (1999). Human language technologies evaluation in the european framework. In *DARPA Broadcast News Workshop, ISBN-1-55860-638-6*, p. 237–242, Whashingthon: Morgan Kaufman Publishers.

MARQUEZ & PADRO (1998). On the evaluation and comparison of taggers: the effect of noise in test corpora. In *COLING/ACL*, Montreal.

S. P. D. (1998). The nist role in automatic speech recognition benchmark tests. In *1st International Conference on Language Resources and Evaluation (LREC98)*, volume 1, p. 433–441, Granada, Spain.

SCHWENK H. (1999). Using boosting to improve a hybrid hmm/neural network speech recognizer. In *IEEE International Conference On Acoustics, Speech, and Signal Processing*., Phoenix, USA.

SIMPSON P. K. (1990). *Artificial Neural Systems - Foundations, Paradigms, Applications and Implementations*. Pergamon Press, 1 edition.

SPARCK-JONES K. & GALLIERS J. (1995). *Evaluating Natural Language Processing Systems*. Springer-Verlag.

TUFIS D. (1999). *Tiered Tagging and combined classifier*, In *Text, Speech and Dialogue*. Lecture Notes in Artificial Intelligence 1692, Jelineck F. and Nörth E. eds. Springer.