Annotation dynamique dans le corpus italien de dialogues spontanés LUNA

Christian Raymond † **et Kepa Joseba Rodriguez* ‡

† Laboratoire Informatique d'Avignon, BP1228, 84911 Avignon Cedex 9, France
‡ Piedmont Consortium for Information Systems (CSI-Piemonte). Italie
christian.raymond@univ-avignon.fr kepajoseba.rodriguez@csi.it

ABSTRACT

In the context of the LUNA project, this paper presents the semantic annotation procedure we are following on an italian corpus. This corpus consists in human-human spontaneous dialogues recorded in the call center of the help desk facility of the Consortium for Information Systems of the Piedmont. The aim of our semantic annotation procedure is to speed up and make more reliable the manual annotation of the corpus. This procedure consists in using a statistical learner to annotate automatically at the semantic level transcribed files and to generate automatically annotated files in the input format of the annotation tool : human annotators have just to check and correct these annotations instead of starting from scratch. In order to converge as fast as possible to reliable automatic annotations and so minimizing the human effort, this procedure follows the active learning paradigm. The active learning procedure is coupled with an annotation error detection to assure more reliable annotation.
Keywords: `Active Learning, semantic annotation, errors detection`

1. Introduction

Le but du projet LUNA est de mener des recherches sur le problème de la compréhension de la parole spontanée dans le contexte de systèmes conversationnels impliqués dans des tâches complexes. Trois étapes sont considérées dans le processus de compréhension : la génération d'étiquettes sémantiques, la composition sémantique des ces étiquettes pour obtenir des structures conceptuelles plus complexes et une validation contextuelle de ces structures. Les modules de compréhension seront appris et évalués sur les corpus LUNA et utilisés dans différents systèmes de dialogue en italien, français et polonais.
Dans ce papier, nous présentons la procédure d'annotation sémantique que nous suivons sur un corpus italien. Ce corpus est constitué de dialogues humain-humain enregistrés dans le centre d'appel du service d'assistance technique du Consortium for Information Systems of the Piedmont (CSI). Le but de notre procédure d'annotation est d'accélérer l'annotation manuelle des données tout en contrôlant la qualité des annotations produites. Le premier point trouve une solution dans le principe d'apprentissage actif (Active Learning) qui a été appliqué dans de nombreux domaines avec beaucoup de succès, dont la reconnaissance et compréhension de la parole spontanée [3]. Le second point est abordé par des techniques de détection d'erreurs d'annotation. Ces deux points sont cruciaux lors de l'annotation de données et ont tout intérêt à être abordés conjointement [11, 12, 8].
Notre procédure consiste à utiliser un algorithme statistique pour annoter automatiquement au niveau sémantique des transcriptions. Notre système génère automatiquement des fichiers au format utilisé par le logiciel d'annotation : les annotateurs humains contrôlent et corrigent les erreurs produites par notre système plutôt que de partir de zéro. Dans le but de converger au plus vite vers des annotations fiables et ainsi minimiser l'effort humain, notre procédure suit le principe d'apprentissage actif, elle est également couplée avec des stratégies de détection d'erreurs.
Nous présentons le corpus, données et annotations, dans la section 2. La section 3 présente le principe d'apprentissage actif tandis que la section 4 présente celui de détection d'erreurs. La section 5 décrit notre procédure d'annotation et les premiers résultats sont présentés dans la section 5.4.

2. Corpus

2.1. Description générale des données

Le corpus italien qui est actuellement transcrit et annoté est composé de dialogues spontanés humain-humain enregistrés au centre d'appel du service d'assistance technique du Consortium for Information Systems of the Piedmont (CSI Piedmont [1]). Les sujets principaux des dialogues concernent des problèmes logiciels ou matériels et de leurs issues administratives. Comme ces dialogues sont spontanés, on peut y trouver d'autres sujets mineurs, tels que certaines conversations à propos d'autres personnes ou de vacances, *etc.* Les dialogues enregistrés ont deux participants, l'appelant, un employé public de la région Piedmont, et un opérateur du service d'assistance technique. La structure générale d'un dialogue est la suivante :

1. Un des participants – en général l'opérateur – débute le dialogue.
2. L'opérateur se pressente lui/elle même et demande l'identité de l'appelant.

*Ce travail a été partiellement financé par la commission Européenne - projet LUNA contrat n^o 33549 et par le Marie Curie Excellence Grant pour le projet ADAMACH (contrat n^o 022593).

1. `http://www.csi.it`

3. L'opérateur questionne à propos du problème.
4. L'appelant explique le problème et les deux personnes collaborent pour en trouver les sources.
5. Le problème peut être résolu de différentes manières :
   (a) Les deux participants collaborent pour résoudre le problème.
   (b) L'opérateur résout le problème seul ou donne les instructions à suivre à l'appelant.
6. Les participants clôturent le dialogue.

Comme il est fréquent dans les dialogues spontanés, il y a un grand nombre d'interruptions, de paroles exprimées en même temps, de phrases tronquées ou agrammaticales, quelques statistiques sont disponibles dans le tableau 1.

Dialogues transcrits	180
Temps (min.)	495.29 ($\overline{x} = 2.75$ *min)*
Nombre de tours	9074 ($\overline{x} = 50$ *tours)*
Nombre de mots	66290 ($\overline{x} = 368$ *mots)*
Nombre de mots différents	4715
Nombre de segments annot.	17462 ($\overline{x} = 97$ *segments*)

Table 1: Description du corpus

Dans le projet LUNA, différents niveaux d'annotations sémantiques ont été définis (voir [9]). Dans ce travail la procédure d'annotation a été expérimentée au niveau de l'annotation sémantique en concepts. Ce niveau correspond à l'identification des briques de bases nécessaires à la compréhension complète d'une intervention humaine. Ces briques sont des unités de sens élémentaires désignées sous le nom de concepts et représentés en paires attribut/valeur.

2.2. Annotation sémantique en concepts

Après analyse d'un ensemble de dialogues, nous avons définis 55 concepts ainsi que les contraintes pour les valeurs possibles. Cette représentation a été utilisée pour construire le dictionnaire de concepts utilisé pour l'annotation.

Quelques unes des catégories principales d'annotation sont :
- Applications logicielles
- Composants matériels
- Composants réseaux
- Personnes : Prénoms et Nom de Famille, catégories professionnelles
- Actions pertinentes pour identifier ou résoudre le problème
- Type de documents utilisés
- Codes d'identification d'ordinateurs et documents
- Lieux : institutions et compagnies, adresses, sites web, numéros de téléphone, *etc.*
- Expressions temporelles

Ce dictionnaire est utilisé pour l'annotation et enrichi au fur et à mesure selon les besoins, voir un exemple [2] d'annotation dans le tableau 2. L'outil utilisé est Semantizer [2], qui a été précédemment utilisé pour l'annotation du corpus MEDIA.

3. Apprentissage actif

Le principe d'apprentissage actif (Active Learning) est de sélectionner pour l'annotation les exemples les plus informatifs, et ainsi réduire le nombre d'exemples d'apprentissages nécessaires à un système pour obtenir un niveau de performance donné. Nous utilisons une méthode d'apprentissage actif basé sur l'incertitude [6] qui sélectionne pour annotation les exemples pour lesquels l'algorithme d'apprentissage est le moins confiant. Pour appliquer cette méthode, nous devons avoir à notre disposition un algorithme d'apprentissage ainsi qu'une mesure de confiance associée. Le choix de l'un ou de l'autre n'est pas crucial, toutefois dans notre situation où nous devons annoter des transcriptions manuelles : nous n'avons pas de contraintes de temps-réel ni le besoin d'être robuste aux erreurs de reconnaissance. Les algorithmes discriminants dans cette situation sont performants et ont l'avantage de pouvoir intégrer de multiples sources de connaissances. Outre ces avantages, les Champs Conditionnels Aléatoires (CRF) [5] fournissent la probabilité conditionnelle d'une annotation complète étant donnée les observations et cette probabilité peut être exploitée en tant que mesure de confiance sur l'incertitude de l'annotation [8]. Nous utilisons dans ces travaux une implémentation [4] des CRF.

Opérateur : sto guardando [lex=filler] l' [avete aperta]$_{c1}$ [stamattina]$_{c2}$
<concept1 action :open>
<concept2 temp-partOfDay :morning>
Appelant : sí
Opérateur : [undici]$_{c3}$ [trentanove]$_{c4}$
<concept3 number-cardinal :11>
<concept4 number-cardinal :39>
Appelant : se vuole [la mia RWS]$_{c5}$ [tredici ottocentotrentacinque]$_{c6}$ forse
<concept5 code-typ :rws>
<concept6 code-value :13835>
Opérateur : vediamo se
Appelant : te l' [ho aperta]$_{c7}$ io
<concept7 action :open>
Opérateur : siete [fuori rete]$_{c8}$ proprio
<concept8 problem :off_line>

Table 2: Exemple d'annotation sémantique au niveau conceptuel

4. Détection d'erreur d'annotation

L'idée derrière le principe de détection d'erreur est de ré-annoter les exemples utilisés pour l'apprentissage du modèle au moyen de ce même modèle. Si les annotations automatiques diffèrent ou reçoivent une faible confiance de la part du modèle, les exemples sont potentiellement mal annotés ou bien difficiles à apprendre. Dans [1] ils sortent les exemples d'apprentissage ayant reçus le poids le plus important par l'algorithme de boosting, dans [7] ils utilisent le poids assigné par leur classifieur SVM. Dans [8] la probabilité conditionnelle délivrée par les CRF est utilisée.

5. Annotation active

Nous avons implémenté une approche d'annotation active (figure 1) dans le but de réduire l'effort humain nécessaire pour l'annotation. Cette approche se base sur des méthodes statistiques pour pré-annoter automatiquement les données et ainsi faciliter et réduire l'effort des annotateurs humains. Comme il est précisé dans la section 2.2, l'annotation

2. Traduction : **O :** *Je regarde [filler] l'avez vous ouvert ce matin ?* // **C :** *oui* // **O :** *11 39* // **C :** *si vous voulez mon RWS-ID 13 835* // **O :** *voyons si* // **C :** *je l'ai ouvert* // **O :** *vous êtes déconnecté*

concerne l'identification de concepts représentés sous la forme attribut/valeur. Les méthodes automatiques employées pour l'annotation automatique sont présentées dans les deux sections suivantes.

5.1. Extraction des attributs

Le modèle μ est un CRF qui est utilisé pour faire de la segmentation et de la classification en concept. Nous ramenons le problème à un étiquetage de séquence en associant à chaque mot le concept dont il fait parti ainsi qu'une information (B ou I, pour begin et inside) indiquant sa position à l'intérieur du concept [3] :

concepts :vDC-B vDC-I vDC-I null null null null p-n-B
mots : cento sessanta quattro okay a nome di Angela

μ est appris en utilisant un traditionnel graphe de dépendance du premier ordre. Les paramètres sont les mots ainsi que leur catégorie d'appartenance associés à leur position dans une fenêtre [-3, 2] autour de la position de décision. Les catégories identifiées sont actuellement MOIS, JOUR, NOMBRE et ORDINAL et utiles pour donner plus de pouvoir de généralisation au modèle. Comme nous annotons la transcription produite manuellement, des informations sur les règles de transcription ont été introduites : *e.g.* si le mot a sa première lettre capitalisée ou toutes ses lettres capitalisées.

5.2. Extraction des valeurs

Le modèle μ fournit la segmentation de l'intervention avec chaque segment classifié avec l'attribut du concept qui lui est attribué. Pour produire/déduire la valeur normalisée de ce concept nous utilisons deux techniques selon le type de concept. Pour les concepts où les valeurs possibles sont potentiellement infinies (*e.g.* les nombres) ou potentiellement non observées dans les dialogues déjà étiquetés (*e.g.* les dates) l'extraction de valeur se fait en appliquant des grammaires permettant de générer la liste exhaustive des valeurs possibles. Pour les autres concepts, l'extraction de valeur est assurée par un classifieur. De cette manière aucune supervision supplémentaire n'est nécessaire. Les nouvelles valeurs introduites lors d'un tour de notre procédure seront prises en compte dès le tour suivant. Dans ces travaux le classifieur choisi est BoosTexter [10] une implémentation de l'algorithme de boosting.

5.3. Procédure d'annotation

Dans notre cas, un exemple est un dialogue composé de tours de parole transcrits de l'utilisateur et de l'opérateur du système. Nous commençons avec N dialogues annotés manuellement (étape 1) pour apprendre un premier modèle μ, dans chaque tour de la procédure, les dialogues non annotés manuellement S_U sont automatiquement annotés (étape 2a) au moyen de μ. k dialogues pour lesquels le modèle est le moins confiant vis à vis de son annotation sont sélectionnés et fournis dans le format utilisé par le logiciel d'annotation, Semantizer. Puis S_k est présenté aux annotateurs humains pour contrôle et correction.

3. vDC=valoreDiCodiche & p-n=persona-nome

1. Apprendre un modèle μ avec N dialogues annotés manuellement (S_L)
2. Tant que (annotateurs/données disponibles)
   (a) Utiliser μ pour annoter la partie non-annotée du corpus (S_U) et produire des fichiers Semantizer
   (b) Ordonner automatiquement (S_U) en accord avec la mesure de confiance donnée par μ
   (c) Sélectionner les k dialogues ayant les scores les plus faibles (S_k)
   (d) Demander une supervision (contrôle/correction) humaine sur S_k
   (e) **Utiliser μ pour annoter S_L et produire S_L^a**
   (f) **Regarder les différences entre S_L et S_L^a**
      i. Exemple difficile à apprendre : **Ajouter nouveaux paramètres dans μ**
      ii. Ambigüités d'annotation : **Demander à un annotateur de désambigüiser S_L**
   (g) $S_L = S_L + S_k$
   (h) Apprendre un nouveau modèle μ avec S_L

Figure 1: Procédure d'annotation active : la partie en clair correspond à l'algorithme traditionnel d'apprentissage actif, la partie en gras correspond à la stratégie liée à la détection d'erreurs d'annotation

Les dialogues manuellement supervisés sont ensuite ajoutés à l'ensemble des données d'apprentissage. Un nouveau modèle μ est appris et le processus est répété.

Dans le même temps, μ ré-annote les dialogues utilisés lors de son apprentissage et les différences entre annotations automatiques et manuelles permettent d'exhiber à chaque tour les erreurs ou ambigüités d'annotations, voir partie en gras dans la figure 1.

Un modèle discriminant comme les CRF a tendance à faire du sur-apprentissage, on peut raisonnablement s'attendre à ce qu'il soit capable de reproduire les annotations manuelles sur le corpus d'apprentissage. Dans notre procédure, nous comparons les annotations automatiques produites et les annotations manuelles sur les exemples du corpus d'apprentissage. Si les annotations diffèrent, nous considérons deux cas :

1. l'annotation manuelle est correcte : nous avons affaire à un exemple difficile à apprendre. Si l'exemple est difficile, c'est parce que le modèle ne possède pas les informations nécessaires pour pouvoir discriminer la bonne annotation d'une mauvaise. Dans certains cas, le paramètre à ajouter est très intuitif : par exemple, nous avons observé que le modèle faisait de nombreuses confusions entre deux concepts <application-software> qui peut être porté par les mots « Lotus Notes », « Outlook », *etc.* et <person-name> qui peut être porté par les mots « Roberto », « Marco », *etc.* À l'origine tout ces mots étaient groupés dans une même catégorie (*i.e.* première_lettre_capitalisé) et le contexte local n'est visiblement pas discriminant. Une rapide observation montre qu'un grand nombre de noms italiens finissent par les lettres « a,i,o,e », un nouveau paramètre prenant en compte cette information a été ajouté et a rendu notre modèle plus performant.

2. l'annotation manuelle n'est pas correcte : corriger l'annotation.

		Annotation automatique							
				segmentation+classification			attribut/valeur		
Act. tour	taille app. en dialog.	# dia.	# tours	tours erronés	taux err. entités	err. *vs.* toutes	tours erronés	taux err. entités	err. *vs.* toutes
1	10	10	561	62.2%	71.2%	1531/2151	62.4%	73.9%	873/1181
2	20	10	517	51.5%	59.5%	1090/1831	55.9%	81.0%	790/975
3	30	10	490	44.5%	54.0%	866/1605	50.2%	72.0%	633/879
4	40	20	1378	43.7%	51.0%	2224/4359	47.5%	71.9%	1635/2274
5	60	20	1547	39.4%	45.7%	2164/4732	48.9%	77.7%	1932/2487
6	80	20	1257	35.0%	42.4%	1497/3533	44.6%	77.3%	1369/1771
7	100	40	2915	32.7%	37.5%	3076/8204	40.5%	63.1%	2674/4237
8	140	40	2103	27.7%	34.5%	1865/5398	33.0%	53.6%	1467/2736

Table 3: Statistiques sur l'annotation active en termes de segmentation/classification et attribut/valeur pour chaque tour d'annotation active

5.4. Évaluation

Nous avons évalué les performances de l'annotation automatique. Après chaque tour d'annotation active, nous comparons les annotations automatiques et finales (*i.e.* les annotations corrigées par les annotateurs humains). Nous mesurons la faculté de la procédure à produire la segmentation/classification et à produire la valeur normalisée pour un attribut. Le tableau 3 présentent les performances de l'annotation automatique. Dans le premier tour (première ligne du tableau), nous utilisons $N = 10$ dialogues annotés manuellement pour apprendre un modèle, nous l'utilisons pour annoter $k = 10$ dialogues transcrits composés de 561 tours de parole. Après correction par les annotateurs, nous comparons les annotations automatiques et manuelles : le modèle produit des annotations contenant 62.2% de tours erronés en termes de segmentation+classification (*i.e.* 62.2% de ces tours durent être corrigés) le taux d'erreur entités[4] est de $\frac{1531}{2151}$ =71.2%, 1531 d'entités erronées sur un total de 2151. En termes d'attribut/valeur, 73.9% de ces entités[5] durent être corrigées. Cela peut être interprété comme un taux d'erreur élevé, toutefois, malgré un corpus d'apprentissage très réduit (10 dialogues), un tiers des tours ont été correctement annotés. Le pourcentage d'annotations à corriger diminue à chaque tour d'annotation et dans le dernier tour seulement un tiers d'entre elles ont du être corrigées, à l'exception des valeurs normalisées où 53.6% d'entre elles ont du être corrigées ; ce faible résultat est expliqué par le grand nombre de valeurs possibles, a peu près 1500.

Le premier objectif dans cette expérience, réduire la réduction du temps d'annotation nécessaire, a été achevé. Au début de la procédure les annotateurs utilisaient entre 80 et 90 minutes pour annoter un dialogue, après le troisième tour d'annotation active seulement 25 à 35 minutes furent nécessaires pour corriger les annotations automatiques. Dans les tours suivants le temps nécessaire est resté constant.

6. Conclusion

Nous avons présenté une procédure d'annotation active qui annote automatiquement des dialogues humain-humain au niveau conceptuel permettant de simplifier le travail des annotateurs humains qui ne doivent plus que les corriger. Cette procédure se base sur le principe d'apprentissage actif afin de converger au plus vite vers une annotation automatique robuste et intègre une stratégie basée sur la détection d'erreurs d'annotation. Les premiers résultats en montre l'intérêt : les corrections faites par les annotateurs diminuent au fur et à mesure de la procédure ainsi que leur temps de travail.

4. ici couple séquence de mots+attribut
5. ici couple attribut+valeur

Références

[1] S. Abney, R. Schapire, and Y. Singer. Boosting applied to tagging and PP attachment. In *EMNLP/VLC*, 1999.

[2] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. Semantic annotation of the french media dialog corpus. In *InterSpeech*, 2005.

[3] D. Hakkani-Tür, G. Riccardi, and G. Tur. An active approach to spoken language processing. *ACM Trans. Speech Lang. Process.*, 3(3) :1–31, 2006.

[4] Taku Kudo. Crf++. Disponible en ligne à `http://crfpp.sourceforge.net/`.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[6] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *ICML*, 1994.

[7] T. Nakagawa and Y. Matsumoto. Detecting errors in corpora using support vector machines. In *ACL*, 2002.

[8] C. Raymond and G. Riccardi. Learning with noisy supervision for spoken language understanding. In *ICASSP*, 2008.

[9] C. Raymond, G. Riccardi, K.J. Rodriguez, and J. Wiśniewska. Luna corpus : an annotation scheme for a multi-domain multi-lingual dialogue corpus. In *DECALOG*, 2007.

[10] R. Schapire and Y. Singer. BoosTexter : A boosting-based system for text categorization. *Machine Learning*, 39 :135–168, 2000.

[11] G. Tur, M. Rahim, and D. Hakkani-Tür. Active labeling for spoken language understanding. In *EUROSPEECH*, 2003.

[12] A. Vlachos. Active Annotation. In *EACL*, 2006.