

Atelier ETAPE - Tâche Entité-Nommée

Frédéric Béchet, Benoît Favre

Université d'Aix-Marseille (AMU)
Laboratoire d'Informatique Fondamentale - LIF/CNRS



Rennes, 18 et 19 juin 2012



Contexte

- Participation à ESTER 1 et ESTER 2
 - système **LIANE** librement disponible
 - voir ma page perso sur www.lif.univ-mrs.fr
- chaîne d'analyse linguistique MACAON
 - étiquetage morphosyntaxique, chunking, analyse en dépendance
 - analyse de graphes de mots, production de graphes d'hypothèses
- modèles linguistiques spécifiques à l'oral
 - projet ANR DECODA
 - prise en compte des spécificités de l'oral spontané dans la chaîne d'analyse
 - disfluences, agrammaticalités, énoncés tronqués
- Entités nommées et résolution d'entités
 - "entity linking" dans les projets *EdyLex* et *PERCOL/REPERE*



Développement d'un système LIF pour ETAPE-EN

- De grands projets !!
 - décodages joints POS/CHUNK/EN/PARSING dans MACAON
 - prise en compte de l'oral avec les modèles DECODA
 - intégration avec la RAP via des graphes de mots
 - résolution d'entités pour une optimisation globale
- hélas
 - manque de temps, de ressources humaines
 - ça devient difficile de participer à des projets sans financement !!
 - complexité du modèle d'annotation des EN
 - difficile de capitaliser sur ESTER2
- Résultat
 - système peu ambitieux de style **YACT** (*Yet Another CRF Tagger*)
 - intégrant tout de même certaines briques originales
 - limitant les dégâts au niveau de l'évaluation



Le système LIF-ETAPE-EN

- Combinaison de modules existant pour produire les paramètres décrivant les données
- Paramètres syntaxiques
 - MACAON “oral” pour les POS
 - Analyseur en constituant (Berkeley parser - PTB)
- Paramètres sémantiques
 - Base de connaissance ALEDA (Sagot et al. XX)
 - Lexique d'hyperonyme (projet EdyLex)
 - Système Entités Nommées d'ESTER 2 - LIANE



Le système LIF-ETAPE-EN

- Méthodes d'apprentissage pour la prédiction des étiquettes EN
 - Etiquetage de séquences - CRF (*CRFsuite*)
 - Prédiction d'étiquettes - classifieur à base de boosting (*IcsiBoost*)
 - Apprentissage sur les corpus QUAERO et ETAPE
- Production des annotations EN structurées
 - simplification du modèle
 - système entièrement dirigé par les données
 - aucune règle, rien de spécifique aux annotations QUAERO



Simplification du modèle d'annotation

Comment utiliser un CRF linéaire pour faire de la prédiction structurée ?

- Structure des EN
 - deux types d'étiquettes :
 - les composants : name, kind, ...
 - les entités : pers, loc, ...
 - enchassement des entités sur plusieurs niveaux
- Exemple

mots	composants	Niveau 4	Niveau 3	Niveau 2	Niveau 1
membre du cabinet de l' ancien ministre de la défense	B.kind B.kind B.qualifier B.kind B.name	 B.org.adm	 B.func.ind l.func.ind l.func.ind l.func.ind	 B.org.ent l.org.ent l.org.ent l.org.ent l.org.ent l.org.ent l.org.ent	B.func.ind l.func.ind l.func.ind l.func.ind l.func.ind l.func.ind l.func.ind l.func.ind



Simplification du modèle d'annotation

- Nombre de mots annotés en entités nommées
 - ESTER-QUAERO = 216K/1,2M (17,3%)
 - ETAPE train+dev = 44K/442K (10,1)
- Répartition des étiquettes selon les niveaux d'entités

mesures	ESTER-QUAERO	ETAPE train+dev
Nb de mots	1,2M	442K
Nb de mots \in entité	216K	44K
Nb de mots \in entité niveau 1	186K (86,1%)	40K (87%)
Nb de mots \in entité niveau 2	29K (13,6%)	5,7K (12,8%)
Nb de mots \in entité niveau 3	572 (0,3%)	100 (0,2%)
Nb de mots \in entité niveau 4	17 (0,008%)	1 (0,00002%)

- Simplification : On peut se limiter à deux niveaux d'enchassement



Simplification du modèle d'annotation

- Modèle d'annotation sur 2 niveaux d'entités + composants
 - niveau entités 1 & 2 → étiquetage de séquences (*CRFsuite*)
 - composantss → classification connaissant l'étiquetage en entités (*IcsiBoost*)
- Exemple

mots	composant	Niveau 2	Niveau 1
le			
Premier ministre	B.kind		B.func.ind
Driss	B.name.first		B.pers.ind
Jettou	B.name.last		I.pers.ind
a			
fait			
hier	B.name		B.time.date.rel
une			
déclaration			
sur			
l'			
action			
du			
gouvernement	B.name		B.org.adm
durant	B.time-modifier		B.time.date.rel
les	I.time-modifier		I.time.date.rel
8	B.val	B.amount	I.time.date.rel
derniers	B.qualifier	I.amount	I.time.date.rel
mois	B.unit	I.amount	I.time.date.rel



Simplification du modèle d'annotation

- Modèle d'annotation sur 2 niveaux d'entités + composants
 - niveau entités 1 & 2 → étiquetage de séquences (*CRFsuite*)
 - composants → classification connaissant l'étiquetage en entités (*IcsiBoost*)
- Exemple

mots	composants	Niveau 2	Niveau 1
le			
Premier ministre	B.kind	B.func.ind	B.func.ind
Driss	B.name.first	B.pers.ind	B.pers.ind
Jettou	B.name.last	I.pers.ind	I.pers.ind
a	NULL	NULL	NULL
fait	NULL	NULL	NULL
hier	B.name	B.time.date.rel	B.time.date.rel
une	NULL	NULL	NULL
déclaration	NULL	NULL	NULL
sur	NULL	NULL	NULL
l'	NULL	NULL	NULL
action	NULL	NULL	
du	NULL	NULL	NULL
gouvernement	B.name	B.org.adm	B.org.adm
durant	B.time-modifier	B.time.date.rel	B.time.date.rel
les	I.time-modifier	I.time.date.rel	I.time.date.rel
8	B.val	B.amount	I.time.date.rel
derniers	B.qualifier	I.amount	I.time.date.rel
mois	B.unit	I.amount	I.time.date.rel



Apprentissage des modèles

- Développement du système sur QUAERO + ETAPE
 - Apprentissage : QUAERO + 46 fichiers ETAPE
 - Développement : 5 fichiers ETAPE
 - Test : 8 fichiers ETAPE
- Choix pour l'apprentissage des CRF pour les niveaux d'entités 1 & 2
 - prédiction indépendante, séquentielle ou jointe

Prédiction	Rappel	Précision	F-mesure
indépendante Niveau 1 (<i>indep1</i>)	69,4	85,9	76,8
indépendante Niveau 2 (<i>indep2</i>)	68,0	83,4	74,9
Séquentielle Niveau 1&2 (<i>seq1&2</i>) (avec Niveau 1 = ref)	97,6	97,5	97,6
<i>indep1+indep2</i>	70,8	79,6	75,0
<i>indep1+seq1&2</i>	74,2	86,3	79,8
<i>joint 1&2</i>	64,9	81,9	72,4

- Temps d'apprentissage limité (environ 2h par CRF)



Apprentissage des modèles

- Classifieur pour rajouter les composants (une fois l'étiquetage en entité réalisé)
 - paramètres : mot + POS + étiquettes d'entités de niveau 1 & 2
 - étiquettes à prédire :

award-cat , century , day , demonym , demonym.nickname , extractor
kind , month , name.first , name.last , object , qualifier , range-mark
reference-era , time-modifier , title , unit , val , week , year , zip-code

- performance (2000 itérations de boosting)
 - taux d'erreur de classification avec les entités de référence
 - train=7,5% developpement=8,9% test=9,2%



Chaîne de traitement

- texte → tokenizer = tokens
- tokens → MACAON = POS tags
- tokens → LIANE = ESTER NE tags
- POS tags → Berkeley parser = paramètres syntaxiques
- tokens → ALEDA = NAME tags
- tokens → Edylex = hyperonyms
- tokens + POS tags + ESTER NE tags + par. synt. + NAME tags + hyperonyms → CRF Niveau 1 = NE Niveau1
- tokens + POS tags + ESTER NE tags + par. synt. + NAME tags + hyperonyms + NE Niveau 1 → CRF Niveau 2 = NE Niveau2
- tokens + POS tags + NE Niveau1 + NE Niveau2 → IcsiBoost = Composants
- token + Modifier + NE Niveau1 + NE Niveau2 → Aligner texte = soumission ETAPE



Traitement des sorties RAP

- Ne jamais se fier à la casse et aux ponctuations des sorties de RAP !!
- On projette toute les données vers une forme décapitalisée, sans ponctuation
- On reapprend tous les modèles
- Même traitement sur les données de test



Résultats

EST2BC FRE FR 20100208 1750 FINTER DEBATE	SER=59.2	FMES=54.3
EST2BC FRE FR 20101018 0910 FINTER DEBATE	SER=49.5	FMES=59.0
EST2BC FRE FR 20101014 2152 FINTER DEBATE	SER=53.4	FMES=60.9
EST2BC FRE FR 20101007 2152 FINTER DEBATE	SER=65.0	FMES=50.4
EST2BC FRE FR 20100208 1000 FINTER DEBATE	SER=50.6	FMES=59.8
LCP TopQuestions 2011-05-18 000400	SER=44.5	FMES=56.6
LCP CaVousRegarde 2011-05-12 235900	SER=35.7	FMES=67.7
TV8 LaPlaceDuVillage 2011-05-12 172800	SER=49.9	FMES=54.8
BFMTV BFMStory 2011-05-31 175900	SER=39.9	FMES=62.7
LCP PileEtFace 2011-05-26 192800	SER=37.2	FMES=63.5
LCP TopQuestions 2011-05-25 213800	SER=42.8	FMES=59.5
LCP EntreLesLignes 2011-05-06 192800	SER=42.0	FMES=60.6
EST2BC FRE FR 20101024 2004 FINTER DEBATE	SER=43.6	FMES=59.3
LCP EntreLesLignes 2011-05-13 192800	SER=38.0	FMES=58.2
TV8 LaPlaceDuVillage 2011-05-03 201300	SER=56.1	FMES=48.0



Conclusions

- Maintenant on va pouvoir s'amuser avec le corpus !!
- De grands projets !!
 - décodages joints POS/CHUNK/EN/PARSING dans MACAON
 - prise en compte de l'oral avec les modèles DECODA
 - intégration avec la RAP via des graphes de mots
 - résolution d'entités pour une optimisation globale

