Active Learning for Natural Language Parsing and Information Extraction,

de Cynthia A. Thompson, Mary Elaine Califf et Raymond J. Mooney

Philippe Gambette

http://philippe.gambette.free.fr/SCOL/ActiveLearningForNLP.pdf

Introduction

- En ***traitement automatique des langues naturelles*** :
  - beaucoup de données ***non étiquetées*** disponibles
  - ***coût de l'étiquetage*** important

Introduction

- En ***traitement automatique des langues naturelles*** :
  - beaucoup de données ***non étiquetées*** disponibles
  - ***coût de l'étiquetage*** important

  ↳ étiqueter le minimum de données, seulement celles utiles pour l'apprentissage

Introduction

- En ***traitement automatique des langues naturelles*** :
  - beaucoup de données ***non étiquetées*** disponibles
  - ***coût de l'étiquetage*** important

  ↳ étiqueter le minimum de données, seulement celles utiles pour l'apprentissage

  ↳ ***apprentissage actif*** !

Introduction

- En ***traitement automatique des langues naturelles*** :
  - beaucoup de données ***non étiquetées*** disponibles
  - ***coût de l'étiquetage*** important

  ↳ étiqueter le minimum de données, seulement celles utiles pour l'apprentissage

  ↳ ***apprentissage actif*** !

- Tâches ***plus complexes*** qu'un simple étiquetage :
  - apprentissage de règles de ***parsing sémantique***
  - ***extraction d'informations***

Apprentissage actif

Initialisation :

Ensemble d'exemples non étiquetés : U

Ensemble d'exemples étiquetés : $L = \emptyset$

Etiquetage de n exemples (passent de U à L)

Boucle d'ajout d'exemples :

Tant que NONSTOP,

    Entraîner le classifieur sur L,

    Pour tout exemple x de U :

        Etiqueter x avec le classifieur entraîné sur L,

        Calculer l'incertitude de cette étiquette,

    Choisir les k exemples avec la plus grande incertitude

Apprentissage actif

Initialisation :

Ensemble d'exemples non étiquetés : U

Ensemble d'exemples étiquetés : $L = \emptyset$

Etiquetage de n exemples (passent de U à L)

Boucle d'ajout d'exemples :

Tant que NONSTOP,

    Entraîner le classifieur sur L,

    Pour tout exemple x de U :

        Etiqueter x avec le classifieur entraîné sur L,

        Calculer l'incertitude de cette étiquette,

    Choisir les k exemples avec la plus grande incertitude

Formule à choisir, adaptée au type de problème !

Paramètres à régler !

Apprentissage actif

Initialisation :

Ensemble d'exemples non étiquetés : U

Ensemble d'exemples étiquetés : $L = \emptyset$

Etiquetage de n exemples (passent de U à L)

Boucle d'ajout d'exemples :

Tant que NONSTOP,

Entraîner le classifieur sur L,

Pour tout exemple x de U :

Etiqueter x avec le classifieur entraîné sur L,

Calculer l'incertitude de cette étiquette,

Choisir les k exemples avec la plus grande incertitude

VARIANTES

- choix de l'exemple minimisant l'erreur attendue

Apprentissage actif

Initialisation :

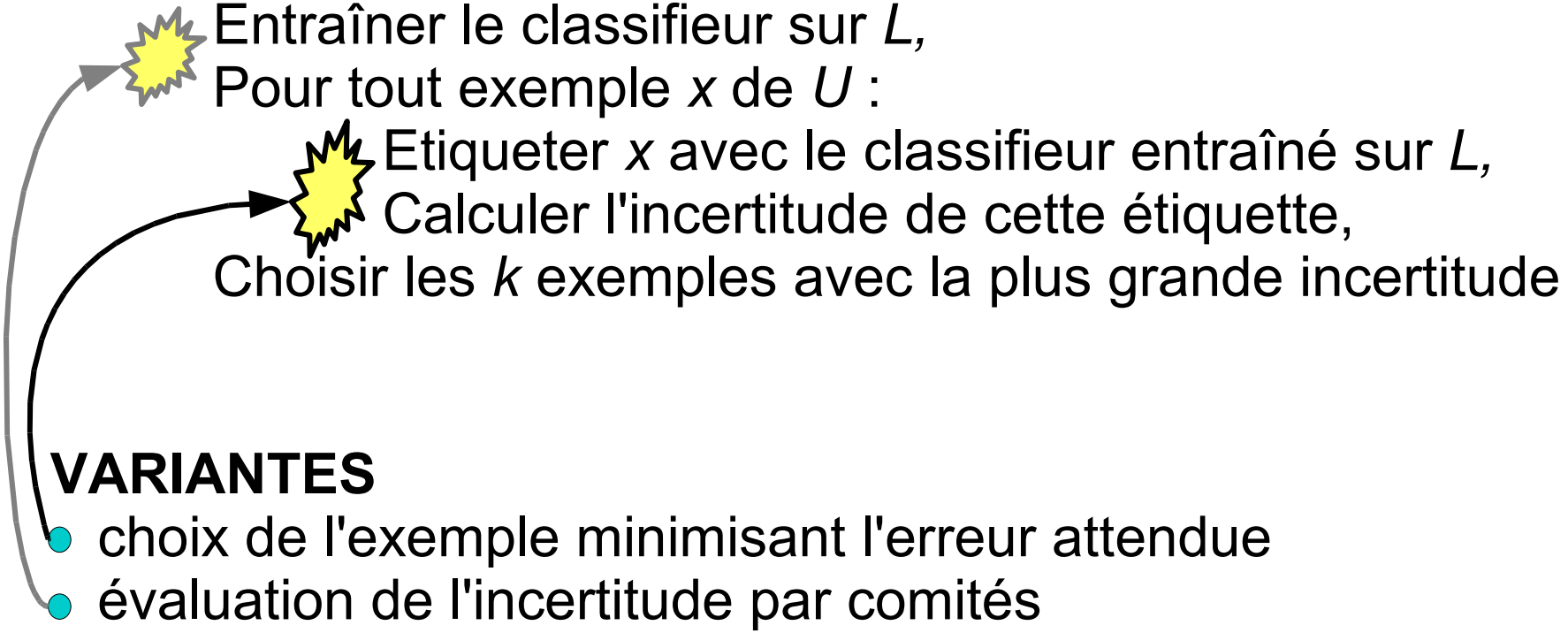
Ensemble d'exemples non étiquetés : U

Ensemble d'exemples étiquetés : $L = \emptyset$

Etiquetage de n exemples (passent de U à L)

Boucle d'ajout d'exemples :

Tant que NONSTOP,



Entraîner le classifieur sur L,

Pour tout exemple x de U :

Etiqueter x avec le classifieur entraîné sur L,

Calculer l'incertitude de cette étiquette,

Choisir les k exemples avec la plus grande incertitude

VARIANTES

- choix de l'exemple minimisant l'erreur attendue
- évaluation de l'incertitude par comités

CHILL

Exemples de couples **(phrase,parsing sémantique)**

("What is the capital of Texas ?"**,**answer(A, (capital(B,A), equal(B,stateid(texas)))**)**

```
1.  ps([answer(freevar,freevar):[]],                shift
       [what,is,the,capital,of,texas,?])
2.  ps([answer(freevar,freevar):[what]],            shift
       [is,the,capital,of,texas,?])
3.  ps([answer(freevar,freevar):[is,what]],         shift
        [the,capital,of,texas,?])
4.  ps([answer(freevar,freevar):[the,is,what]],     introduce
       [capital,of,texas,?])
5.  ps([capital(freevar,feevar):[],
        answer(freevar,freevar):[the,is,what]],     co-reference
       [capital,of,texas,?])
6.  ps([capital(freevar,pvar(0)):[],
        answer(pvar(0),freevar):[the,is,what]],     shift
       [capital,of,texas,?])
7.  ps([capital(freevar,pvar(0)):[capital],
        answer(pvar(0),freevar):[the,is,what]],     shift
       [of,texas,?])
```

Création pour chaque étape de parsing d'exemples positifs et négatifs

Apprentissage de règles par l'algorithme d'induction CHILLIN

Welcome to Geoquery!

A Learning Natural-Language Interface to a US Geography Database

http://www.cs.utexas.edu/users/ml/geo-demo.html

Parser
(ensemble de règles)

Thèse de John Zelle : *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers,* (1995)

CHILL + Active Learning

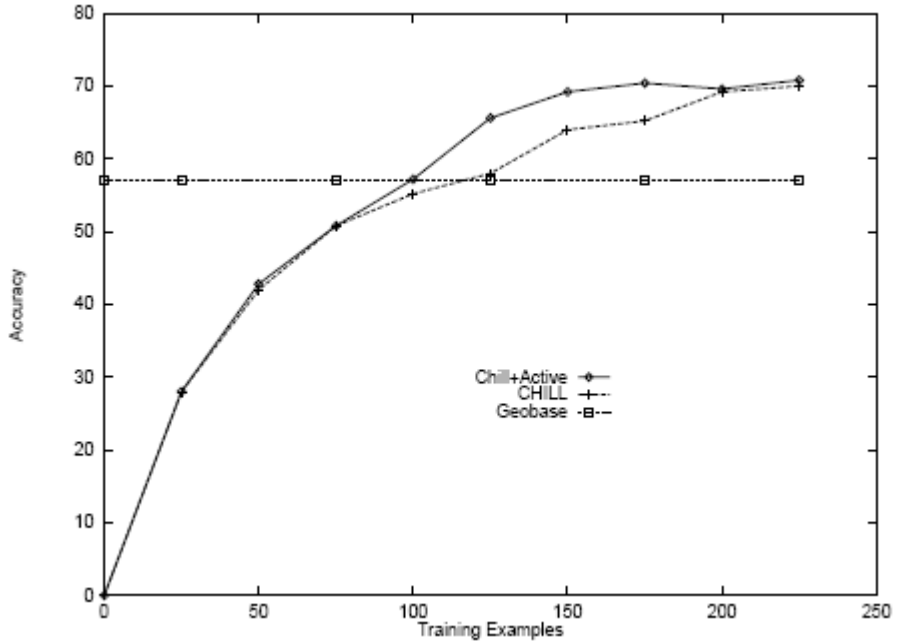
Certitude du **parsing sémantique** d'une phrase ?

- La phrase n'a **pas pu être parsée** :
  exemple très **incertain**
  certitude = **« profondeur » de parsing**
- La phrase a pu être parsée :
  certitude = **moyenne** sur toutes les règles utilisées
  de la **certitude de chaque règle**.
  règle **incertaine** = vérifiée par **peu d'exemples**.

CHILL + Active Learning

Evaluation de l'algorithme ?

- Comparer ***CHILL + Active Learning*** à ***CHILL + random***
- Efficacité :
  - domaine de 250 questions parsées par un expert
  - efficacité = pourcentage de réponses correctes après requête à la base de données

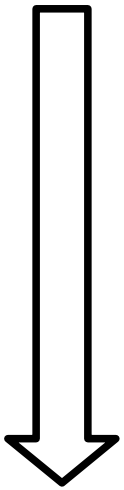


- Apprentissage des règles meilleur que l'expert
- Bon niveau d'efficacité atteint avec moins d'exemples
- Efficacité tout de même limitée...

RAPIER

Identifier dans des offres d'emplois **plusieurs champs prédéfinis** : salaire, ville, type d'emploi → **formulaire**

Exemples de couples **(annonce,formulaire)**



Apprentissage de règles par un algorithme d'**apprentissage relationnel de bas en haut :** trouver des règles **de plus en plus générales**

Ensemble de règles : **motifs**

3 blocs : *termes avant* + *cible* + *termes après*
avec pour chacun des 3 blocs,
contraintes sémantiques ou de syntaxe

Mary Elaine Califf, Raymond J. Mooney : *Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction,* (2003)

RAPIER + Active Learning

Remarque : utilisation d'un **algorithme incrémental**

Certitude d'un formulaire ?

- Somme des certitudes de chaque champ.
- Certitude d'un **champ** :

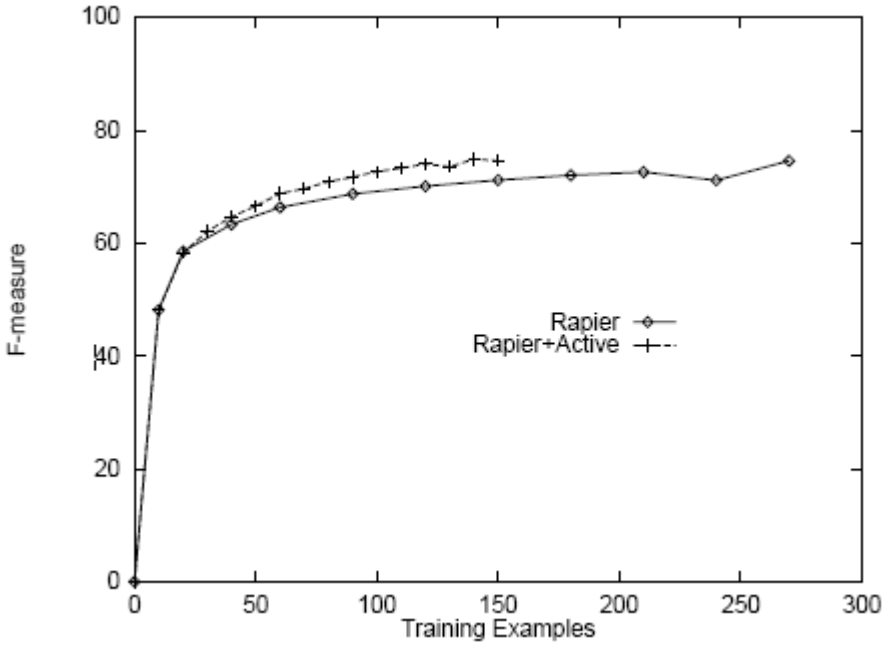
  - Soit le champ est **vide** :
  incertitude liée au **taux de remplissage du champ**
  (champ souvent rempli → très incertain)
  - Soit il est **plein** :
  certitude = certitude de la **règle appliquée**
  (ou min si plusieurs valeurs du champ)
  certitude d'une règle = *nombre d'exemples positifs* – 5 *nombre d'exemples négatifs pour la règle*

Quelle formule exactement ??

RAPIER + Active Learning

Evaluation de l'algorithme ?

- Comparer ***RAPIER + Active*** à ***RAPIER + random***
- Efficacité :
  - 300 posts annotés, validation croisée à 10 plis
  - **bruit** et **couverture** sur tous les champs
  - efficacité = F-measure



- Bon niveau d'efficacité atteint avec moins d'exemples
- Efficacité un peu meilleure qu'avec le problème précédent, mais toujours pas ahurissante

Mon avis

- Article clair, apparemment pas ambigu. Bons exemples (indispensables), **exemples de règles** ?

- Apport de l'**apprentissage actif** :
50 % d'exemples en moins, **pas très impressionnant**

- Quelques mystères d'implémentation :
  - **condition d'arrêt** ?
  - influence du **choix des paramètres** ?
  - **temps de calcul** ?

- Passage à l'échelle et utilité :
  - parsing sémantique de corpus uniquement **spécialisés**,
  - intérêt devant la **théorie sens-texte** ?

Quelles étaient les règles trouvées automatiquement mieux que l'expert pour GeoBase?