

Atelier Etape

Participation du LI à la tâche EN

LI / Université François Rabelais Tours

18 juin 2012



① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

① Participation du LI

- 3 systèmes
- CasEN
- Adaptation à Etape

② Résultats

- En chiffres
- Adjudication ...

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

En chiffres

Adjudication ...

Plan

Plan

① Participation du LI

3 systèmes

CasEN

Adaptation à Etape

② Résultats

Participation à Etape

3 systèmes différents

- **CasEN** \Rightarrow Approche symbolique
- **mXS** \Rightarrow Approche fouille de texte
- **Hybride mXS / CasEN**

Participation à Etape

3 systèmes différents

- **CasEN** \Rightarrow Approche symbolique
- **mXS** \Rightarrow Approche fouille de texte
- **Hybride mXS / CasEN**

CasEN / CasSys

- **CasEN** = ressource
 - Ensemble de grammaires à passer en cascade sur un texte
 - Grammaires intégrant plusieurs niveaux linguistiques simultanément
 - Recherche des motifs les plus certains pour diminuer l'ambiguïté et réduire l'espace de recherche des suivants
- **CasSys** = outil pour utiliser des grammaires en cascade
 - Système permettant de passer une cascade de transducteurs (Abney, 1996)
 - Intégré à l'interface Unitex, utilisable en ligne de commande (sous linux/windows)

CasEN / CasSys

- **CasEN** = ressource
 - Ensemble de grammaires à passer en cascade sur un texte
 - Grammaires intégrant plusieurs niveaux linguistiques simultanément
 - Recherche des motifs les plus certains pour diminuer l'ambiguïté et réduire l'espace de recherche des suivants
- **CasSys** = outil pour utiliser des grammaires en cascade
 - Système permettant de passer une cascade de transducteurs (Abney, 1996)
 - Intégré à l'interface Unitex, utilisable en ligne de commande (sous linux/windows)

CasEN / CasSys

- **CasEN** = ressource
 - Ensemble de grammaires à passer en cascade sur un texte
 - Grammaires intégrant plusieurs niveaux linguistiques simultanément
 - Recherche des motifs les plus certains pour diminuer l'ambiguïté et réduire l'espace de recherche des suivants
- **CasSys** = outil pour utiliser des grammaires en cascade
 - Système permettant de passer une cascade de transducteurs (Abney, 1996)
 - Intégré à l'interface Unitex, utilisable en ligne de commande (sous linux/windows)

Formalisme des transducteurs

- Permet d'exprimer les phénomènes linguistiques complexes en utilisant plusieurs niveaux d'analyse simultanément
 - Niveau lexical
 - Filtres morphologiques
 - Utilisation des traits (grammaticaux, sémantiques)
placés par les dictionnaires sur chaque mot
 - Interdiction de contextes
- Fusionner/remplacer les motifs trouvés par l'entrée des transducteurs avec sa sortie

Formalisme des transducteurs

- Permet d'exprimer les phénomènes linguistiques complexes en utilisant plusieurs niveaux d'analyse simultanément
 - Niveau lexical
 - Filtres morphologiques
 - Utilisation des traits (grammaticaux, sémantiques)
placés par les dictionnaires sur chaque mot
 - Interdiction de contextes
- Fusionner/remplacer les motifs trouvés par l'entrée des transducteurs avec sa sortie

Formalisme des transducteurs

- Permet d'exprimer les phénomènes linguistiques complexes en utilisant plusieurs niveaux d'analyse simultanément
 - Niveau lexical
 - Filtres morphologiques
 - Utilisation des traits (grammaticaux, sémantiques)
placés par les dictionnaires sur chaque mot
 - Interdiction de contextes
- Fusionner/remplacer les motifs trouvés par l'entrée des transducteurs avec sa sortie

Formalisme des transducteurs

- Permet d'exprimer les phénomènes linguistiques complexes en utilisant plusieurs niveaux d'analyse simultanément
 - Niveau lexical
 - Filtres morphologiques
 - Utilisation des traits (grammaticaux, sémantiques)
placés par les dictionnaires sur chaque mot
 - Interdiction de contextes
- Fusionner/remplacer les motifs trouvés par l'entrée des transducteurs avec sa sortie

Adaptation à Etape

Pour Ester2

- Grammaires de pers, org, loc
- Début de travail sur func, amount, prod, time

Pour Etape

- 1ère phase
 - Adaptation à l'annotation Quaero
 - Mise en place des composants
- 2ème phase
 - Prise en compte des disfluences de l'oral
 - Amélioration des grammaires : func.ind, amount, prod, time
 - Création de grammaires : pers.coll, pers.ind non nommés, func.coll
 - Amélioration des dictionnaires

Adaptation à Etape

Pour Ester2

- Grammaires de pers, org, loc
- Début de travail sur func, amount, prod, time

Pour Etape

- 1ère phase
 - Adaptation à l'annotation Quaero
 - Mise en place des composants
- 2ème phase
 - Prise en compte des disfluences de l'oral
 - Amélioration des grammaires : func.ind, amount, prod, time
 - Création de grammaires : pers.coll, pers.ind non nommés, func.coll
 - Amélioration des dictionnaires

Adaptation à Etape

Pour Ester2

- Grammaires de pers, org, loc
- Début de travail sur func, amount, prod, time

Pour Etape

- 1ère phase
 - Adaptation à l'annotation Quaero
 - Mise en place des composants
- 2ème phase
 - Prise en compte des disfluences de l'oral
 - Amélioration des grammaires : func.ind, amount, prod, time
 - Création de grammaires : pers.coll, pers.ind non nommés, func.coll
 - Amélioration des dictionnaires

Adaptation à Etape

Pour Ester2

- Grammaires de pers, org, loc
- Début de travail sur func, amount, prod, time

Pour Etape

- 1ère phase
 - Adaptation à l'annotation Quaero
 - Mise en place des composants
- 2ème phase
 - Prise en compte des disfluences de l'oral
 - Amélioration des grammaires : func.ind, amount, prod, time
 - Création de grammaires : pers.coll, pers.ind non nommés, func.coll
 - Amélioration des dictionnaires

Adaptation à Etape en chiffres ...

Amélioration des dictionnaires utilisés

205000 entrées

- 14% prénoms, 4.5% célébrités, 16% ethnonymes / adjectifs toponymiques
- 60% toponymes (dont 97% villes, 1.25% pays, 0.2% territoires, 0.2% édifices etc.)
- 15% organisations/entreprises/associations
- 15% ergonymes (dont 50% de médias, 15% monnaies, 4% produits etc.)
- 16% professions etc.

Dictionnaire de désambiguation

- Adapté au corpus : 500 mots ambigus
ex : Manuel/ville/prénom, Elysée/batiment/prénom

Grammaires

- cascade de 100 transducteurs
 - Environ 600 sous-graphes
 - 1,6 Mo de grammaires

Adaptation à Etape en chiffres ...

Amélioration des dictionnaires utilisés

205000 entrées

- 14% prénoms, 4.5% célébrités, 16% ethnonymes / adjectifs toponymiques
- 60% toponymes (dont 97% villes, 1.25% pays, 0.2% territoires, 0.2% édifices etc.)
- 15% organisations/entreprises/associations
- 15% ergonymes (dont 50% de médias, 15% monnaies, 4% produits etc.)
- 16% professions etc.

Dictionnaire de désambiguation

- Adapté au corpus : 500 mots ambigus
ex : Manuel/ville/prénom, Elysée/batiment/prénom

Grammaires

- cascade de 100 transducteurs
 - Environ 600 sous-graphes
 - 1,6 Mo de grammaires

Adaptation à Etape en chiffres ...

Amélioration des dictionnaires utilisés

205000 entrées

- 14% prénoms, 4.5% célébrités, 16% ethnonymes / adjectifs toponymiques
- 60% toponymes (dont 97% villes, 1.25% pays, 0.2% territoires, 0.2% édifices etc.)
- 15% organisations/entreprises/associations
- 15% ergonymes (dont 50% de médias, 15% monnaies, 4% produits etc.)
- 16% professions etc.

Dictionnaire de désambiguation

- Adapté au corpus : 500 mots ambigus
ex : Manuel/ville/prénom, Elysée/batiment/prénom

Grammaires

- cascade de 100 transducteurs
 - Environ 600 sous-graphes
 - 1,6 Mo de grammaires

Adaptation à Etape en chiffres ...

Amélioration des dictionnaires utilisés

205000 entrées

- 14% prénoms, 4.5% célébrités, 16% ethnonymes / adjectifs toponymiques
- 60% toponymes (dont 97% villes, 1.25% pays, 0.2% territoires, 0.2% édifices etc.)
- 15% organisations/entreprises/associations
- 15% ergonymes (dont 50% de médias, 15% monnaies, 4% produits etc.)
- 16% professions etc.

Dictionnaire de désambiguation

- Adapté au corpus : 500 mots ambigus
ex : Manuel/ville/prénom, Elysée/batiment/prénom

Grammaires

- cascade de 100 transducteurs
 - Environ 600 sous-graphes
 - 1,6 Mo de grammaires

Adaptation à Etape

- Coût en temps pour corriger/bâtir les grammaires
- Pas d'adaptation spécifique aux tâches asr !

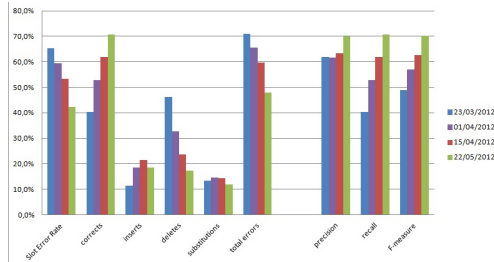


Figure: Evolution SER du dev

Difficulté des allers/retours guide/ref/scoring

- à cause des erreurs de la référence !

Adaptation à Etape

- Coût en temps pour corriger/bâtir les grammaires
- Pas d'adaptation spécifique aux tâches asr !

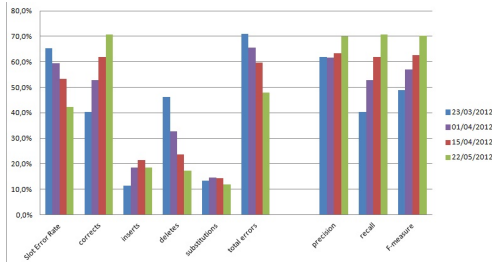


Figure: Evolution SER du dev

Difficulté des allers/retours guide/ref/scoring

- à cause des erreurs de la référence !

Adaptation à Etape

- Coût en temps pour corriger/bâtir les grammaires
- Pas d'adaptation spécifique aux tâches asr !

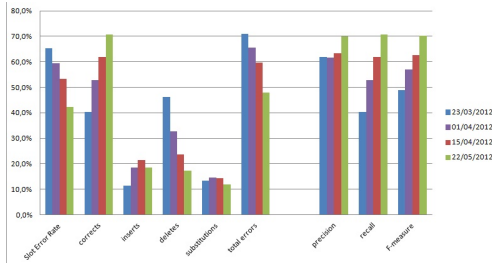


Figure: Evolution SER du dev

Difficulté des allers/retours guide/ref/scoring

- à cause des erreurs de la référence !

① Participation du LI

② Résultats

En chiffres

Adjudication ...

Adaptation à Etape

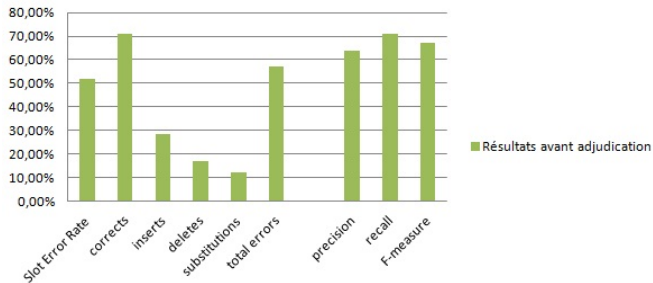


Figure: Résultats avant adjudication

- Beaucoup d'insertions (> 25 %)
- Erreur : les parenthèses (perte de 0,5% de SER)

Adaptation à Etape

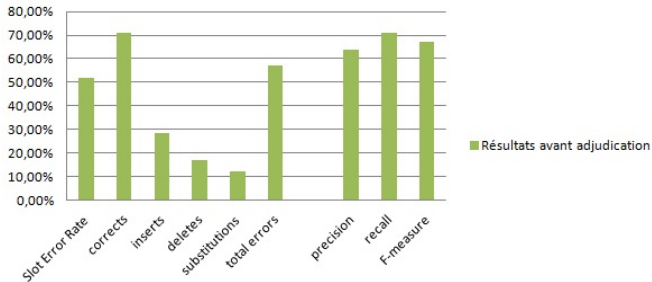


Figure: Résultats avant adjudication

- Beaucoup d'insertions (> 25 %)
- Erreur : les parenthèses (perte de 0,5% de SER)

Résultats

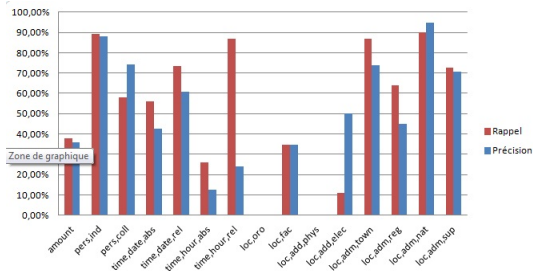


Figure: Résultats avant adjudication

- Très fort rappel/précision : pers.ind, loc.adm.nat, prod.award, prod.fin
- Corrects : loc.adm.town, loc.adm.sup
- Problèmes de précision/rappel : time (insertions : "maintenant", saisons ...), amount (erreur d'interprétation du guide : "quelques émissions" ...)

Résultats

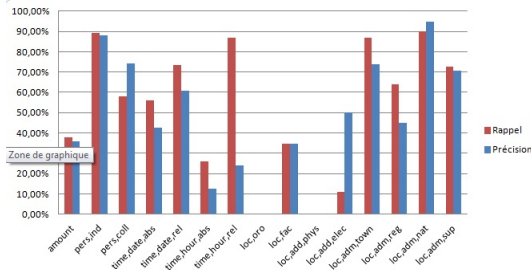


Figure: Résultats avant adjudication

- Très fort rappel/précision : pers.ind, loc.adm.nat, prod.award, prod.fin
- Corrects : loc.adm.town, loc.adm.sup
- Problèmes de précision/rappel : time (insertions : "maintenant", saisons ...), amount (erreur d'interprétation du guide : "quelques émissions" ...)

Résultats

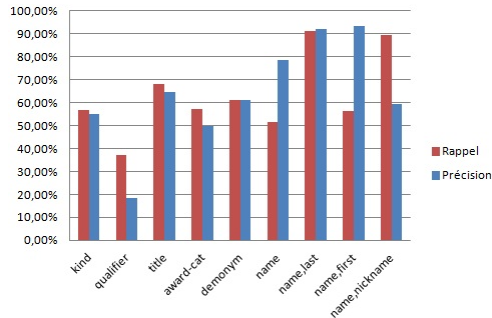


Figure: Résultats avant adjudication

- Très bons résultats : name.last
- Mauvais rappel : name.first (bcp hors-contextes), name, qualifier ...

Résultats

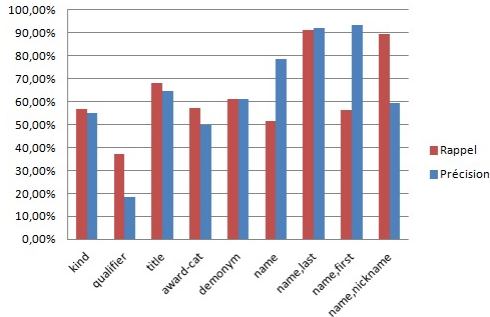


Figure: Résultats avant adjudication

- Très bons résultats : name.last
- Mauvais rappel : name.first (bcp hors-contextes), name, qualifier ...

14/16

Résultats

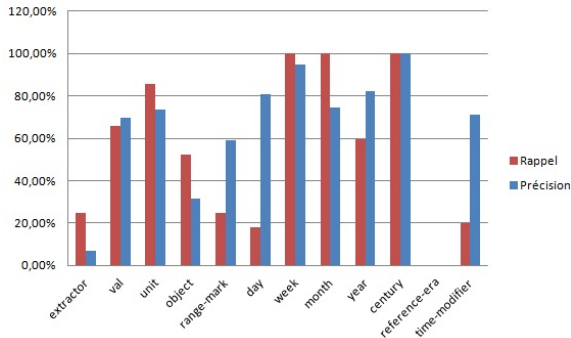


Figure: Résultats avant adjudication

Résultats

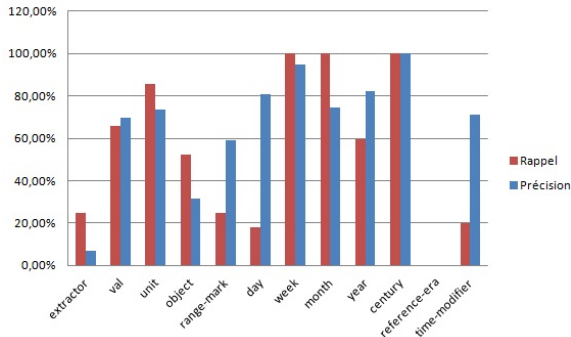


Figure: Résultats avant adjudication

Adjudication ...

Test sur le fichier BFM avec nos propres modifications de la référence !

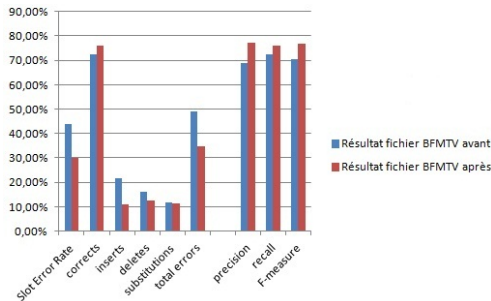


Figure: Scoring ref BFMTV corrige / BFMTV CasEN

- SER final entre 29 et 44 % sur ce fichier
- Insertions divisées par 2 (précision augmentée dans les time)
- Deletions diminuées de 20% environ

Adjudication ...

Test sur le fichier BFM avec nos propres modifications de la référence !

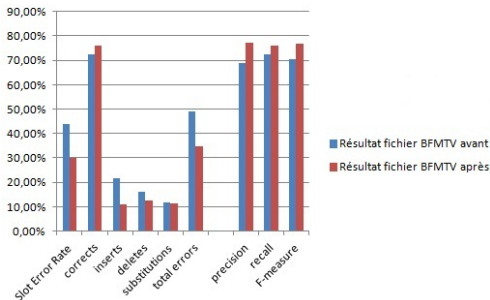


Figure: Scoring ref BFMTV corrige / BFMTV CasEN

- SER final entre 29 et 44 % sur ce fichier
- Insertions divisées par 2 (précision augmentée dans les time)
- Deletions diminuées de 20% environ