Approches à base de fréquences pour la simplification lexicale

A.-L. Ligozat[1,2] C. Grouin[1,3] A. Garcia-Fernandez[4] D. Bernhard[5]

[1]LIMSI, CNRS

[2]ENSIIE

[3]INSERM U872 Eq20 & UPMC

[4]LAS, CNRS/EHESS/Collège de France

[5]LiLPa, Université de Strasbourg

TALN 2013

Plan

Simplification de textes

Objectif

- Rendre les textes plus faciles à lire

$\rightarrow$ enfants, locuteurs non natifs...

Deux sous-tâches

- Simplification syntaxique
  Notre TGV, qui est pourtant parti avec 5 minutes de retard de Montparnasse, est arrivé à l'heure. $\rightarrow$ *Notre TGV est arrivé à l'heure.*
- Simplification lexicale
  exploiter $\rightarrow$ *utiliser*

Simplification lexicale

- Consiste à remplacer des mots ou groupes de mots par des équivalents plus simples
  - Identifier les mots équivalents en contexte
  - Choisir le mot le plus simple

Participation à la tâche de simplification lexicale de l'anglais de SemEval 2012

Tâche

- Mot cible choisi
- Plusieurs substituts possibles à ordonner

Exemple

<instance id="270">
<context>With the growing demand for these fine garden furnishings , they found it necessary to dedicate a portion of their business to <head>outdoor</head> living and patio furnishings .</context>
</instance>

Substituts proposés : *{alfresco, outside, open-air, outdoor}*
Sortie attendue : *(outdoor, open-air, {outside, alfresco})*

Participation à la tâche de simplification lexicale de l'anglais de SemEval 2012

Corpus

Textes courts en anglais issus de documents internet

- Corpus d'apprentissage
  - 300 instances
- Corpus de test (évaluation)
  - 1710 instances

Annotation

Par des locuteurs non natifs de l'anglais

Plan

Caractérisation des éléments lexicaux en contexte

Facteurs de choix

- Critères concernant l'élément lui-même, principalement issus des mesures de lisibilité de textes

mot court	nombre de caractères ou de syllabes
mot courant	fréquence en corpus, présence dans des listes de mots simples
mot compréhensible par des enfants	caractéristiques psycholinguistiques : caractère concret, âge d'acquisition...

  - *véloce* $\Rightarrow$ *rapide*

Caractérisation des éléments lexicaux en contexte

Facteurs de choix

- Critères concernant l'élément lui-même, principalement issus des mesures de lisibilité de textes

mot court	nombre de caractères ou de syllabes
mot courant	fréquence en corpus, présence dans des listes de mots simples
mot compréhensible par des enfants	caractéristiques psycholinguistiques : caractère concret, âge d'acquisition...

  - *véloce* $\Rightarrow$ *rapide*

Caractérisation des éléments lexicaux en contexte

Facteurs de choix

- Contexte local de l'élément

appartenance à un terme ou une collocation	fréquence du n-gram, présence dans des listes de collocations. . .

  - *prononcer un discours* ⇒ *faire un discours*
  - *prononcer trois mots* ⇒ *dire trois mots*

- Contexte plus général

contexte thématique	co-occurrents, LSA. . .
répétition	anaphore, fréquence dans le texte

  - *The film shows Afghan mercenaries to be involved with the separatists (...) . : film* ⇒ *documentary* (*film, movie, picture*).

Caractérisation des éléments lexicaux en contexte

Facteurs de choix

- Contexte local de l'élément

  appartenance à un terme ou une collocation — fréquence du n-gram, présence dans des listes de collocations...

  - *prononcer un discours* ⇒ *faire un discours*
  - *prononcer trois mots* ⇒ *dire trois mots*

- Contexte plus général

  contexte thématique — co-occurrents, LSA...

  répétition — anaphore, fréquence dans le texte

  - *The film shows Afghan mercenaries to be involved with the separatists (...) . : film* ⇒ *documentary* (*film, movie, picture*).

Plan

Méthode 1 : Fréquence des termes

Fondé sur la fréquence des substituts

- Fréquence en corpus
  - Corpus en anglais simplifié : Simple English Wikipedia (SEW)
  - Hypothèse : mots de ce corpus préférés par locuteurs non natifs

Méthode

- Conversion de la SEW au format texte
- Extraction des n-gram de mots ($n = 1$ à 3)
- Calcul des fréquences de n-gram
- Substituts ordonnés par fréquence décroissante

Tests

- Lemmatisation préalable par le TreeTagger
- Variation de la taille des n-gram

Méthode 2 : Contexte local

Tri par fréquence décroissante du mot avec son contexte proche

- Utilisation des n-gram du service Microsoft Web
- Taille du contexte variable : un à quatre token à droite et à gauche du mot cible

Exemple :

He brings an incredibly rich and diverse background that
lush

Méthode 2 : Contexte local

Tri par fréquence décroissante du mot avec son contexte proche

- Utilisation des n-gram du service Microsoft Web
- Taille du contexte variable : un à quatre token à droite et à gauche du mot cible

Exemple : contexte gauche 2 tokens, droite 1 token

He brings an incredibly rich and diverse background that
lush

Méthode 3 : Co-occurrents

Utilisation de listes de co-occurrents

- Ressources de co-occurrences :
  - Wortschatz
  - liste construite à partir de la SEW : mots présents dans une même phrase
- Comparaison du contexte du mot dans la phrase et des co-occurrents

Exemple

Snow covered areas appear <head>bright</head> blue in the image which was taken in early spring and shows deep snow cover .
bright plus fréquemment trouvé en corpus avec *blue* que *vibrant*

Plan

Métrique

Comparaison par paire des listes de rangs fournis par le système avec les rangs de référence

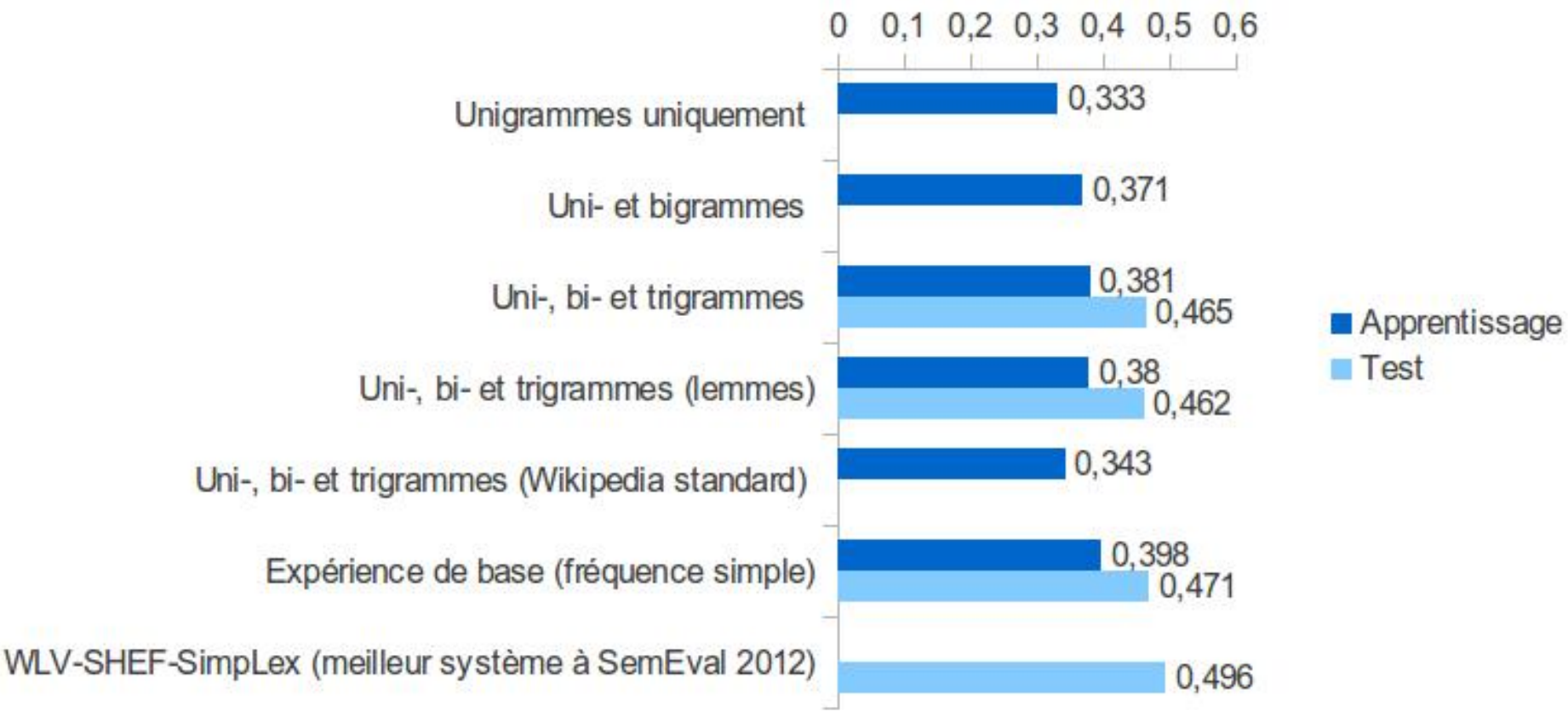
- Comparaison de la position de chaque substitut entre hypothèse et référence
- Coefficient κ d'accord inter-annotateur

Approches baseline

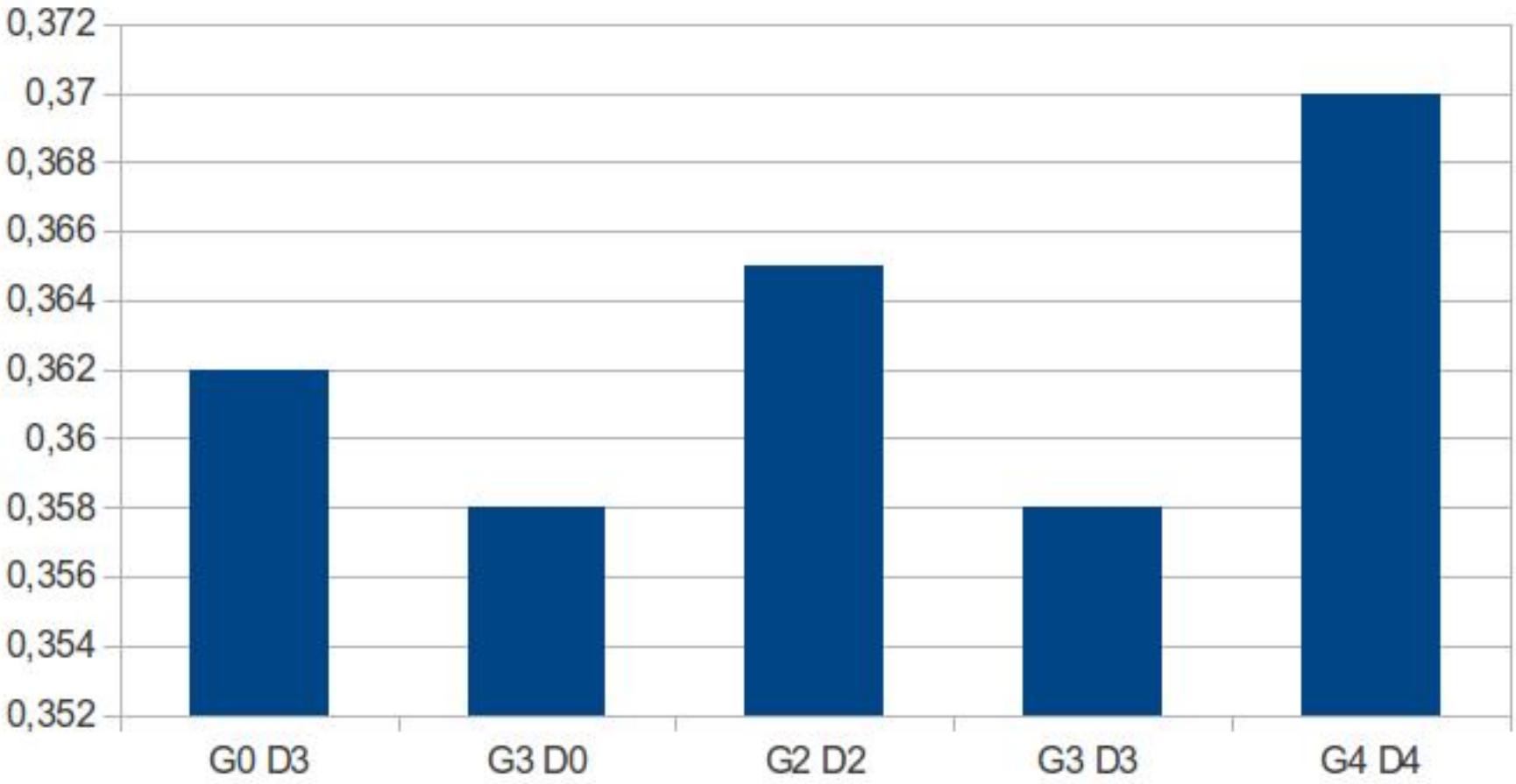
3 baseline fournies par les organisateurs

- Conservation de l'ordre
- Ordonnancement aléatoire
- Fréquence des termes en corpus (Google Web 1T)

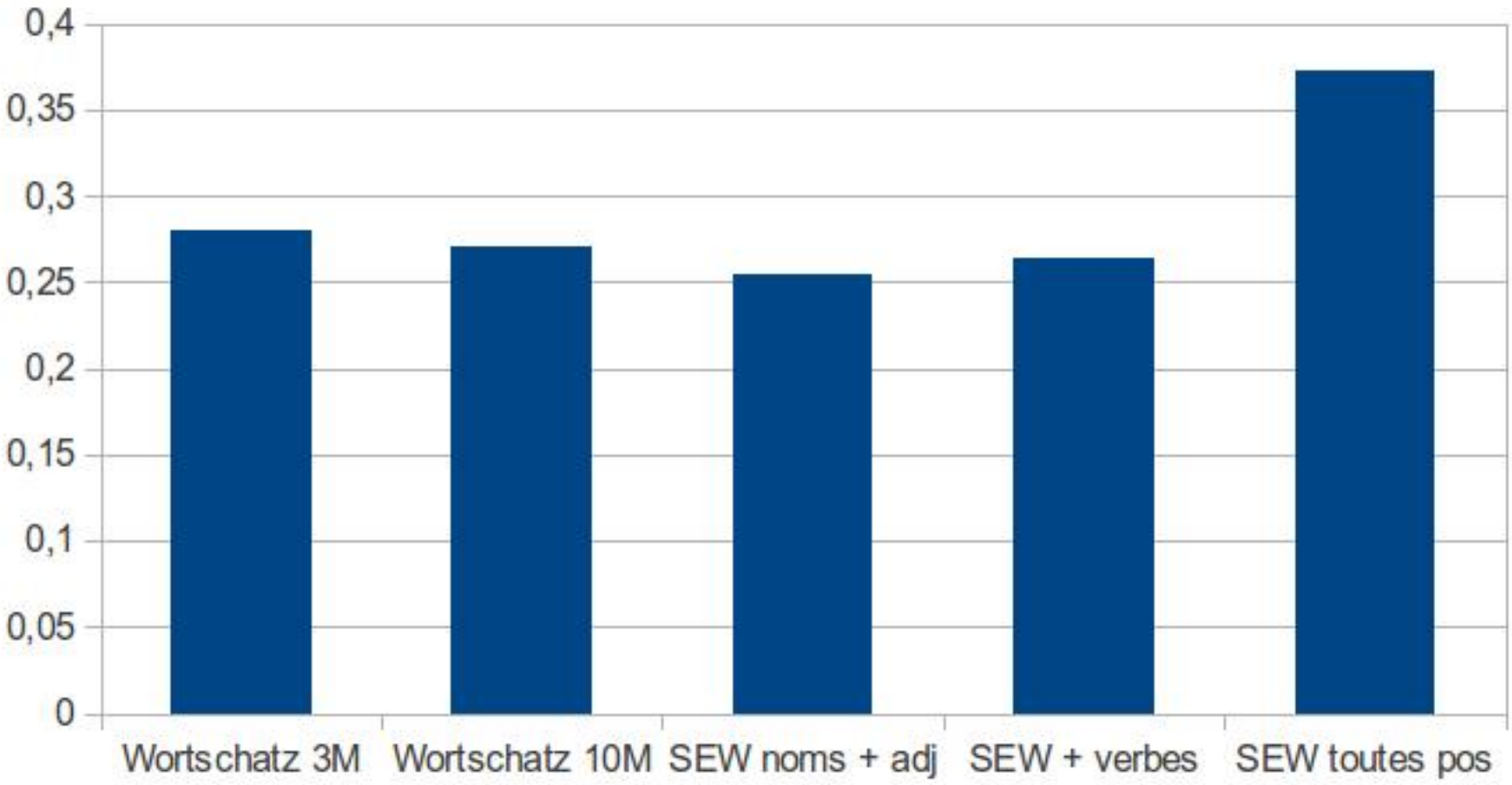
Modèle fondé sur les fréquences des termes



Modèle fondé sur le contexte local



Modèle fondé sur les co-occurrences



Plan

Conclusion

Simplification lexicale

- Tâche difficile, même pour humains
- Meilleur système à peine au-dessus de la baseline “fréquence simple”

Perspectives

- Améliorer la prise en compte du contexte
- Nécessité de construire des corpus pour évaluer cette tâche