SL02353X : Thématiques actuelles de la recherche en TAL & Séminaire de l'axe CARTEL

SemDis tâche 1 : Substitution lexicale par méthodes distributionnelles

Introduction : C. Fabre

Présentation du système BACANAL : B. Gaume

8 décembre 2014

L'atelier SemDis
(enjeux actuels en sémantique distributionnelle)

- Atelier organisé dans le cadre de la conférence TALN 2014
- Equipes organisatrices : CLLE-ERSS et IRIT-Melodi
  - Pour CLLE-ERSS : C. Fabre, N. Hathout, L.-M. Ho-Dac, F. Morlane-Hondère, F. Sajous, L. Tanguy
  - Pour IRIT-MELODI : T. Van de Cruys, P. Muller
- Organisation de 2 tâches :
  - une tâche compétitive de substitution lexicale qui fournit un classement des systèmes en compétition
    - Séance UE TAL 8 décembre
  - une tâche exploratoire sur un corpus spécialisé, le corpus TALN (2007-2013), qui fait l'objet d'un appel à communication ciblé
    - Séance UE TAL 15 décembre (présentation N. Hathout)

La tâche de substitution lexicale

- Etant donné un mot-cible dans une phrase complète, proposer une (ou plusieurs) unités de substitution qui n'altèrent pas le sens global de l'énoncé.
- Le choix du substitut sera confronté aux réponses fournies par des annotateurs humains
- Exemples :
  - Le policier a été surpris par les **feux** nourris d'un groupuscule terroriste => tirs
  - Cette pratique **fonde** sa légitimité dans le cadre de l'État de droit => assure, assoit, garantit…

La tâche de substitution lexicale

- Adaptation au français de la tâche SemEval 2007 *Lexical substitution* (McCarthy & Navigli, 2009)
- Une tâche qui combine 2 opérations :
  - Identifier un ensemble de candidats substituts pour un mot cible
    - Pas de liste prédéfinie. Caque système détermine librement la ressource qu'il utilise.
  - Identifier le meilleur candidat en fonction du contexte dans lequel se trouve le mot cible
- Les données :
  - 2010 phrases (201 mots, 10 phrases par mot)
  - issues du corpus EIC (English Internet Corpus)
  - annotées manuellement par 5 annotateurs (chacun fournissant jusqu'à 3 substituts pour un mot-cible)

Intérêt de la tâche

- Selon (McCarthy & Navigli, 2009) :
  - Une tâche qui peut être utile dans plusieurs applications : simplification de texte, résumé, systèmes de question-réponse, inférence sémantique (détection de paraphrases)
  - Un objectif plus général :

    "a means of examining the issue of word sense representation by giving participants a free reign over the lexical inventories used on a task that evaluates the inventories and also contextual disambiguation." (McC&N)
- Pour nous :
  - Evaluer l'apport d'une ressource distributionnelle pour ce type de tâche
  - Construire un jeu de données réutilisable (Morlane-Hondère et al. 2014)

Les ressources utilisées par les systèmes en 2007

- "The systems all used one or more predefined lexical inventories for obtaining candidate substitutes (…) USYD was the only system to supplement candidates from predefined resources with candidates from corpus data"

Source : (Mc&N 2009)

Table 2 Sources for candidate substitutes

System	WordNet	Macquarie	Roget	Other
MELB (Martinez et al. 2007)	√			
HIT (Zhao et al. 2007)	√			
UNT (Hassan et al. 2007)	√			Encarta
IRST1 (Giuliano et al. 2007)	√			OAWT
IRST2 (Giuliano et al. 2007)	√			OAWT
KU (Yuret 2007)			√	
SWAG1 (Dahl et al. 2007)			√	
SWAG2 (Dahl et al. 2007)			√	
USYD (Hawker 2007)	√	√		Web 1T
TOR (Mohammad et al. 2007)		√		

SemDiS 2014
Le jeu d'évaluation – les mots

- 30 mots cibles : 10 N, 10 V, 10 A
- Mots polysémiques et fréquents
- Critères de choix :
  - présence dans le Robert
  - au moins 2 sens bien identifiables
  - + de 500 occ. dans FRWAC
  - des synonymes nombreux et fréquents

N	V	A
affection, *capacité*, *couverture*, *débit*, *direction*, *don*, *espace*, *intérêt*, *montée*, *vaisseau*	*arrêter, commander, entraîner, éplucher, essuyer, faucher, fonder, interpréter, maintenir, taper*	*aisé, compris, grossier, hermétique, incorrect, mince, modeste, obscur, riche, vaseux*

Le jeu d'évaluation – les phrases

- 300 phrases, 10 phrases pour chaque mot cible
- Utilisation d'un concordancier sur le corpus frwac http://nl.ijs.si/noske/wacs.cgi/first_form
- Chaque phrase doit clairement illustrer un des sens du mot
- Elle doit être bien formée, pas trop longue
- Les différents sens identifiés doivent être illustrés dans les 10

sens	phrase
tuer	La guerre franco-prussienne **faucha** le jeune artiste […]
	Un psychiatre dont le fils a été **fauché** au front croise […]
moissonner	[…] certaines parcelles sont **fauchées** tardivement l'été.
	Sa mission : planter (plus de 2000 arbres), tailler, **faucher**, récolter les fruits, presser les jus pour les propriétaires privés et publics.
voler	On picolait un peu – une bouteille d'alcool **fauchée** chez Ceron.

L'annotation

- Annotateurs francophones, étudiants et chercheurs en linguistique
- 7 annotateurs par phrase
- 3 substituts au maximum
- Extraits des consignes :

  - Votre tâche est de trouver des mots qui peuvent se substituer à ce mot en rouge tout en préservant au maximum le sens de la phrase

  - Vous pourrez proposer jusqu'à 3 substituts, mais si aucun ne vous vient à l'esprit n'insistez pas

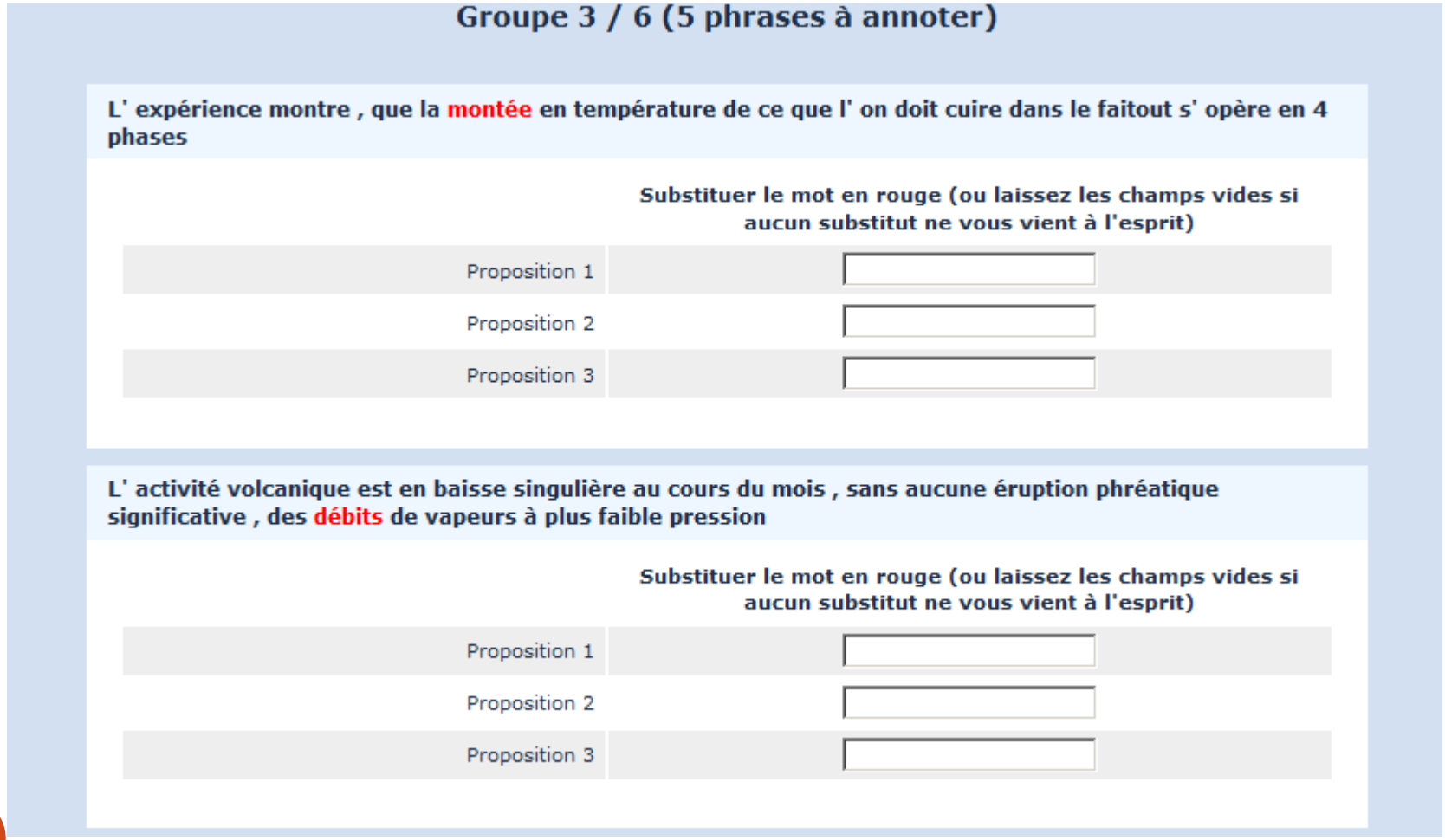
  - Les mots simples sont à privilégier

  - La phrase résultante doit être correcte, mais des modifications syntaxiques légères sont tolérées :

    - Le **gros** garçon s'amuse / Le garçon **obèse**
    - Paul a **échoué** dans sa tentative / Paul a **râté** sa tentative

L'interface d'annotation

- Outil de gestion d'enquêtes LimeSurvey



Résultats

- 4014 substituts, 1734 formes différentes
- Moyenne par phrase :
  - 13 propositions
  - 7 substituts différents
- Exemple :
  - *Cette réorganisation n' [entraînera] aucune suppression de poste et les 69 salariés de l' usine seront transférés sur d'autres sites vauclusiens du groupe.*

    causera (4), provoquera (2), aboutira à (1), causer (1), génèrera (1), amènera à (1), générer (1), amènera (1), impliquer (1), impliquera (1), aura pour conséquence (2)

Nettoyage des données

- 1) Filtrage, normalisation, lemmatisation
  - Validation automatique de 88% des propositions
  - Vérification manuelle des 480 propositions restantes :
    - Suppression des mot outils
    - Lemmatisation des formes ambigües
    - Exclusion de propositions mal formées
  - **Au final : 3961 propositions, 1099 formes différentes**

  Ex : entraîner.v 55 :: causer 5; amener 2; générer 2; avoir pour conséquence 2; provoquer 2; impliquer 2; aboutir 1;
- 2) Traitement des soumissions
  - Verbes pronominaux
  - Infinitif / Adj → ppé

Accord inter-annotateurs

- Accord par paire :
  - Proportion moyenne de réponses identiques pour chaque phrase et chaque paire d'annotateurs
  - 25,8% (27,75% pour la tâche en anglais)
- Accord avec le mode :
  - Proportion moyenne d'annotateurs qui ont inclus dans leurs réponses la réponse la plus fréquente
  - 73% (50,67% en anglais)

Evaluation des systèmes

- Mesures Best et Oot (Mc Carthy & Navigli 2009)
  - Best :
    - Seule la première substitution du système est prise en compte. Le score est élevé si ce substitut a été proposé par un grand nombre d'annotateurs
  - Oot (*out of ten*) :
    - Le score évalue le nombre de réponses des annotateurs couvertes par l'ensemble des propositions du système (max. 10, non ordonnées)
- Baseline :
  - Pour chaque mot-pivot, sélection de tous les synonymes dans DicoSyn (mots simples)
  - Sélection des 10 premiers ordonnés par leur fréquence dans FRWAC

Soumission

- 3 participants :
  - Proxteam (Yann Desalle et al.) : 3 soumissions
  - CEA (Olivier Ferret) : 5 soumissions
  - Alpage (Kata Gábor) : 1 soumission

Le système du CEA LIST (Ferret 2014)

- Génération des substituts :
  - Recherche dans des dictionnaires :
    - Dictionnaire de synonymes de Word XP (*word*)
    - Dictionnaire de synonymes Dicosyn (*isc*)
    - Thésaurus distributionnel FreDist : voisins distributionnels (syntaxiques et cooccurrents) construits à partir d'un corpus de journaux (*fredist*)
- Choix parmi les substituts générés :
  - Mesure de similarité calculée entre chaque candidat substitut et l'ensemble des mots pleins de la phrase
- 5 combinaisons
- Paramètres testés : ressource, mesure de similarité, nature des mots pris en compte dans le choix

Le système d'Alpage WoDis (Gábor 2014)

- Ressources utilisées :
  - WOLF (Sagot et Fišer 2008)

    Candidats substituts = mots du synset + hyperonymes directs
  - Complété par des candidats générés par similarité distributionnelle à partir du corpus FrWiki parsé
- Méthode de désambiguïsation :
  - But : disposer de plus de données pour les sens marginaux (enrichissement des synsets peu fournis par le biais de relations complémentaires)
  - Calcul des contextes spécifiques à chaque synset (à partir du corpus FrWiki)
  - Chaque candidat substitut est ensuite représenté sur cet espace (vecteur de désambiguisation du candidat)
  - Calcul d'une mesure de similarité entre le vecteur de la phrase et le vecteur de désambiguisation de chaque candidat

Résultats globaux

	best	oot
Proxteam_JDM_Syn	**0,097**	**0,402**
CEA_list-word_cos_sent	**0,075**	**0,236**
Proxteam_AxeParaProx_JDM_Syn	0,065	0,357
Alpage_WoDiS	**0,063**	**0,205**
Proxteam_LM	0,051	0,212
baseline_dicosyn	*0,045*	*0,325*
CEA_list-fredist_cos_sent	0,040	0,236
CEA_list-isc_cos_w2	0,037	0,284
CEA_list-isc_cos_sent	0,033	0,287
CEA_list-isc_l2_sent	0,010	0,231

Résultats par catégorie

	best			oot		
	N	A	V	N	A	V
Proxteam_JDM_Syn	0,1102	0,1058	0,0749	0,3977	**0,4285**	0,3787
cea_list-word_cos_sent	0,0753	0,0739	0,0759	0,1947	0,2447	**0,2678**
ProxTeam_AxeParaProx_JDM_Syn	0,0553	0,0540	0,0871	0,3112	**0,3958**	0,3626
Alpage_WoDiS	0,0544	0,0720	0,0613	0,1909	0,2107	**0,2130**
Proxteam_LM	0,0521	0,0400	0,0611	0,2326	0,1663	0,2370
baseline_dicosyn	0,0436	0,0404	0,0520	0,2937	0,3362	0,3435
cea_list-fredist_cos_sent	0,0318	0,0283	0,0599	0,1812	0,2245	0,3026
cea_list-isc_cos_w2	0,0295	0,0408	0,0407	0,2427	0,2810	0,3287
cea_list-isc_cos_sent	0,0253	0,0343	0,0402	0,2329	0,2873	0,3395
cea_list-isc_l2_sent	0,0035	0,0116	0,0147	0,1628	0,2299	0,3000