Analyse d'une tâche de substitution lexicale : quelles sont les sources de difficulté ?

RÉSUMÉ

Nous proposons dans cet article une analyse des résultats de la campagne SemDis 2014 qui proposait une tâche de substitution lexicale en français. Pour les 300 phrases du jeu de test, des annotateurs ont proposé des substituts à un mot cible, permettant ainsi d'établir un *gold standard* sur lequel les systèmes participants ont été évalués. Nous cherchons à identifier les principales caractéristiques des items du jeu de test qui peuvent expliquer les variations de performance pour les humains comme pour les systèmes, en nous basant sur l'accord inter-annotateurs des premiers et les scores de rappel des seconds. Nous montrons que si plusieurs caractéristiques communes sont associées aux deux types de difficulté (rareté du sens dans lequel le mot-cible est employé, fréquence d'emploi du mot-cible), d'autres sont spécifiques aux systèmes (degré de polysémie du mot-cible, complexité syntaxique).

ABSTRACT

Difficulty analysis for a lexical substitution task

This paper provides an analysis of the results of the SemDis 2014 evaluation campaign dedicated to a lexical substitution task in French. A gold standard has been established consisting of a dataset of 300 sentences, each of them associated with a list of substitutes that annotators proposed for a given target word. Our aim is to identify the main characteristics of this dataset that have an impact on human annotation and on the performance of the systems that have participated in the campaign. Our evaluation is based on the inter-annotator agreement scores and on the recall of the systems. We show that while several characteristics are found to have an impact on both aspects (level of rarity of the target word sense, frequency of the word), some are specific to the systems (degree of polysemy of the target word and characteristics pertaining to the sentence context).

MOTS-CLÉS : substitution lexicale, difficulté d'une tâche, annotation.

KEYWORDS: lexical substitution, task difficulty, annotation.

1 Introduction

Les données produites dans le cadre d'une campagne d'évaluation peuvent être exploitées pour identifier les facteurs qui déterminent le niveau de difficulté de la tâche visée et qui influent sur les performances des systèmes. Les objectifs de ce type d'étude sont multiples : prédire la difficulté d'un traitement, augmenter la pertinence des *gold standards* pour mieux prendre en compte les différents paramètres de la tâche, cibler les efforts à produire pour améliorer les performances des systèmes, concevoir des systèmes adaptatifs en fonction des caractéristiques des données en entrée. Cette approche a en particulier été mise en œuvre dans le domaine de la recherche d'information (Carmel & Yom-Tov, 2010), mettant au jour le poids des caractéristiques linguistiques de la requête (taille et complexité syntaxique de la requête, présence de noms propres), ou de la collection ciblée. La tâche de désambiguïsation sémantique (WSD) a également donné lieu à de nombreuses études

visant à comprendre les sources de difficulté pour l'annotation (Passonneau *et al.*, 2010) comme pour les systèmes (Palmer *et al.*, 2007).

Dans cet article, nous adoptons cette même démarche pour l'étude d'une tâche de substitution lexicale en français. Cette tâche, qui associe désambiguïsation et recherche d'équivalent sémantique, met en jeu des processus dont la compréhension peut guider les traitements sémantiques lexicaux en contexte, en particulier ceux mis en oeuvre dans le domaine de la sémantique distributionnelle. Nous avons fait le choix de nous intéresser à la fois aux systèmes participants (*runs*) et aux annotateurs qui ont permis de constituer le gold pour cette tâche. Les deux aspects sont liés puisque la mesure des systèmes dépend en partie de la dispersion des propositions des annotateurs. Néanmoins, ce double point de vue sur la tâche met au jour des caractéristiques complémentaires. Après avoir décrit la tâche, nous présentons les caractéristiques prises en compte et qui concernent à la fois les phrases, les mots à substituer et les sens de ces mots. Nous mettons en évidence les caractéristiques les plus nettement corrélées à la difficulté de la tâche pour les annotateurs et pour les systèmes, de façon globale, puis selon la catégorie grammaticale du mot-cible.

2 Difficulté de la substitution lexicale

La tâche compétitive de substitution lexicale en français proposée dans le cadre de l'atelier SemDis [1] est inspirée de celle proposée initialement dans la campagne SemEval par McCarthy & Navigli (2009). Pour évaluer et classer les systèmes participant à cette tâche, le jeu d'évaluation suivant a été construit :

- 30 mots-cibles : 10 adjectifs, 10 noms et 10 verbes, tous polysémiques et fréquents ;
- 300 phrases : 10 pour chacun des mots-cibles, sélectionnées dans le corpus FR-WAC et illustrant différents sens ou emplois des mots-cibles ;
- un *gold standard* constitué des substituts proposés par 7 juges pour chacune des phrases ; chaque juge pouvait proposer jusqu'à 3 réponses par phrase.

Les substituts proposés et le degré de dispersion des réponses changent selon la phrase, le mot à substituer, et le sens de ce mot. Ainsi, les réponses aux deux phrases suivantes, qui impliquent toutes deux le mot *espace*, s'organisent très différemment.
120 : *J'aime toucher et sentir la matière et ne pas laisser d'espace entre mes mains et ce que je crée*.
229 : *Notre vocation est de mettre en valeur vos espaces extérieurs, leur donner un style qui vous corresponde.*
Dans la première phrase, les 7 juges ont proposé les substituts suivants (le nombre entre parenthèses est le nombre de juges) : *vide* (7), *distance* (3), *place* (2), *interstice* (1), *intervalle* (1), *séparation* (1). Dans la deuxième, les substituts sont : *lieu* (2), *zone* (2), *emplacement* (1), *endroit* (1), *place* (1), *superficie* (1), *environnement* (1). On constate qu'un substitut a reçu le suffrage de tous les juges dans la phrase 120 (*vide*), alors qu'aucun substitut évident ne s'impose dans la phrase 229.

Pour quantifier ce degré d'accord, nous avons utilisé l'entropie normalisée en utilisant la formule standard $H = -\sum_i p_i log(p_i)/log(N)$ où p_i est la probabilité qu'un annotateur propose le substitut numéro i et N est le nombre total de réponses des annotateurs. Cette valeur d'entropie est comprise entre 0 et 1 et est faible lorsque peu de substituts différents ont été proposés, avec un fort regroupement entre annotateurs, et forte lorsqu'aucun accord ou presque n'a été observé. Par exemple, nous obtenons

1. Atelier organisé en 2014 par les équipes CLLE-ERSS et IRIT-Melodi dans le cadre de TALN. Les données d'annotation et d'évaluation sont disponibles librement sur le site de l'atelier : `https://www.irit.fr/semdis2014/fr/`

une entropie de $0,55$ pour la phrase n^{o}120 et de $0,86$ pour la phrase n^{o}229, ce qui reflète bien le fait que la seconde manifeste un accord bien moindre dans les substituts proposés.

Les réponses des systèmes participants sont évaluées par une mesure de rappel fondée sur les propositions des juges, la qualité d'un substitut étant évaluée par le nombre de juges qui l'ont proposé. Si deux mesures sont utilisées dans (Fabre *et al.*, 2014), nous nous concentrons ici sur la mesure *OOT* (*out of ten*) qui totalise, pour les 10 substituts proposés par le système, le nombre d'annotateurs qui ont proposé chacun d'entre eux. Par exemple, pour la phrase 229, un des systèmes a produit : *distance ; aire ; terrain ; zone ; lieu ; surface ; temps ; région ; écart ; étendue*. Seuls les mots *zone* et *lieu* ont été proposés par les annotateurs, chacun deux fois. Ces 2×2=4 points sont normalisés par le score maximal envisageable, à savoir le nombre total de substituts proposés par les annotateurs (ici 9), ce qui donne un OOT de 0,44.

Pour quantifier globalement la difficulté des systèmes automatiques, nous avons calculé le score OOT moyen obtenu par les 9 systèmes (ou versions de systèmes) qui ont concouru. Lorsqu'on observe le calcul du score OOT, il est clair qu'il est en grande partie corrélé avec l'entropie des substituts proposés par les annotateurs : un système ne peut espérer obtenir un score important si les substituts de référence sont trop dispersés, ce que confirme le coefficient de corrélation fortement négatif ($r = -0,62$). Pour isoler le comportement moyen des systèmes en neutralisant le rôle de l'entropie, nous avons donc utilisé pour notre analyse les résidus d'un modèle linéaire calculé en prenant l'entropie comme variable. Autrement dit, les valeurs correspondent à la variation du score OOT moyen qui n'est pas expliquée directement par l'entropie.

Ce sont donc ces deux mesures (entropie des annotations et score OOT moyen résiduel) que nous allons étudier en les mettant en regard des caractéristiques linguistiques des items.

3 Caractéristiques étudiées

Nous avons mesuré pour chaque item du jeu d'évaluation un ensemble de caractéristiques en nous inspirant en particulier des travaux similaires menés en désambiguïsation sémantique. Passonneau *et al.* (2010) ont par exemple montré que des traits relatifs à la fois à la sémantique du mot-cible et aux particularités des contextes phrastiques étaient susceptibles d'affecter le niveau d'accord des annotateurs et les performances des systèmes.

Nous avons regroupé les caractéristiques en trois niveaux : celui du mot-cible, celui du sens que prend le mot-cible dans la phrase, et enfin celui de la phrase elle-même. La plupart des caractéristiques se fondent sur les sorties de l'analyseur Talismane que nous avons utilisé pour segmenter, étiqueter et parser les phrases du jeu de test (Urieli & Tanguy, 2013). Nous avons privilégié des caractéristiques génériques et synthétiques, qui pourront être affinées si cette première analyse met en évidence des liaisons.

Niveau du mot-cible

Pour chacun des 30 mots-cibles, indépendamment de leur acception et de la phrase dans laquelle ils ont été présentés aux juges, nous avons calculé les caractéristiques suivantes :

— **Fréquence du mot-cible** : nous avons estimé la fréquence d'emploi de chaque mot-cible (lemme) telle qu'elle est donnée dans la version 1.2.1 du lexique GLAFF (Sajous *et al.*, 2013), qui se base sur une version étiquetée et lemmatisée du corpus FRWaC. Dans la suite, c'est le

logarithme de cette fréquence qui sera utilisé. La fréquence absolue varie de 759 occurrences (*vaseux*) à 427 900 (*espace*).

— **Nombre de synonymes du mot-cible** : nous avons extrait le nombre de synonymes disponibles pour le mot-cible, indépendamment des différents sens que celui-ci peut recouvrir. Pour cela, nous nous sommes basés sur les renvois analogiques du *Robert* présents dans le dictionnaire DicoSyn (Ploux & Victorri, 1998). Ce nombre de synonymes varie de 5 (*vaseux*) à 60 (*grossier*).

— **Nombre de sens du mot-cible** : cette variable numérique code le nombre de sens distincts du mot-cible représentés dans le jeu de test, comme précisé dans le paragraphe suivant. Le nombre de sens par mot-cible varie de 2 (pour 12 mots-cibles différents) à 5 (pour *couverture*). Notons que le nombre de phrases pour chaque sens n'est pas nécessairement homogène pour un mot donné, bien qu'aucune disproportion importante n'ait été relevée.

Niveau du sens

Pour ce deuxième ensemble de caractéristiques, nous avons tout d'abord identifié, pour chaque mot-cible, le sens dans lequel il était employé dans chaque phrase du jeu de test. Ce travail a été fait manuellement en nous référant aux sens identifiés dans le *Larousse* interrogeable en ligne [2]. Nous avons opté pour ce dictionnaire en raison du choix qu'il opère pour l'ordonnancement des sens : à la différence d'autres dictionnaires comme le *Trésor de la Langue Française* et le *Robert*, les lexicographes du *Larousse* proposent un classement basé principalement sur la fréquence d'emploi estimée (Pruvost, 2006; Larousse, 2005), plaçant les sens usuels avant les sens spécifiques, ce qui constitue une information intéressante à exploiter.

L'attribution des sens a été effectuée par 3 juges (les auteurs de l'article) avec une décision finale prise collectivement en cas de désaccord. Pour faciliter la présentation des données, chaque sens a été étiqueté en employant un terme arbitraire, généralement un synonyme ou un hyperonyme distinctif (comme */riche/* et */facile/* pour distinguer les deux sens de *aisé*).

— **Rang du sens** : il s'agit du rang du sens identifié dans le *Larousse*. Ce rang a été normalisé pour prendre une valeur entre 0 (premier sens dans le dictionnaire, donc le plus fréquent) et 1 (dernier sens, donc le plus rare).

— **Nombre de synonymes du sens** : en alignant les sens identifiés avec ceux présentés dans la ressource extraite de DicoSyn, il nous a été possible d'estimer plus précisément le nombre de synonymes pour chaque sens du mot-cible. Cette valeur varie de 0 (pour une dizaine de cas, comme le sens */diplôme/* de *capacité*) à 43 (pour le sens */énigmatique/* de *obscur*).

Niveau de la phrase

Ce dernier niveau correspond à l'insertion d'un mot-cible (avec un sens précis) dans un contexte phrastique.

— **Fréquence moyenne des mots de la phrase** : nous avons calculé la moyenne des logarithmes des fréquences de chaque nom, verbe, adjectif et adverbe de la phrase, à partir des mêmes ressources de référence que pour le mot-cible. Les scores varient de 2,5 pour la phrase *La prolifération de bactéries et de champignons dans la bouche entraîne une coloration et une transformation des papilles gustatives.* à 6,7 pour la phrase *Ce n'est plus l'État mais les marchés qui commandent l'économie...*

— **Complexité syntaxique** : nous avons mesuré la complexité syntaxique de chaque phrase en

2. http://www.larousse.fr/dictionnaires/francais/

utilisant l'analyse en dépendances comme le font (Tanguy & Tulechki, 2009), nous avons retenu la **distance** moyenne (en nombre de mots) entre deux mots reliés par une relation de dépendance.

— **Position du mot-cible** : nous avons calculé le rang (normalisé) de la position du mot-cible dans la phrase (0 = début de phrase, 1 = fin de phrase).

4 Analyse et résultats

Notre objectif est d'identifier quelles caractéristiques des items (parmi celles listées dans la section précédente) sont les plus susceptibles d'avoir influencé les propositions des annotateurs et des systèmes.

Pour mesurer l'impact des caractéristiques sur l'entropie des annotations d'une part, et sur le score OOT moyen d'autre part, nous avons utilisé le coefficient de corrélation de Pearson. La table 1 présente, pour les deux mesures visées, la liste des facteurs significativement ($p < 0,05$) corrélés avec la difficulté qu'ont eue les annotateurs et les systèmes à traiter les items du jeu de test. Nous avons indiqué la valeur absolue du coefficient de corrélation, le sens de la liaison étant orienté vers une difficulté croissante (entropie élevée et score *OOT* bas).

Mesure	**Facteurs significativement corrélés à la difficulté**
Entropie des annotations	rang du sens (0,32), nombre de synonymes (0,18), fréquence moyenne des mots de la phrase (0,17), fréquence du mot-cible (0,16),
Score *OOT* moyen résiduel	nombre de sens (0,37), nombre de synonymes (0,23), rang du sens (0,18), fréquence du mot-cible (0,12)

TABLE 1 – Facteurs de la difficulté, par r décroissant

Le rang du sens se dégage nettement comme facteur de difficulté pour les annotateurs, qui ont atteint un meilleur accord pour les sens les plus usuels des mots-cibles (selon le *Larousse*). Parmi les cas les plus marqués, notons que pour *arrêter* on voit plus d'accord pour le sens */stopper/* que pour */interpeller/* ou */décider/*. Pour l'adjectif *mince* le sens */fin/* a plus d'accord que */faible/*. Les autres caractéristiques liées au désaccord sont le nombre de synonymes du mot-cible et sa fréquence d'emploi.

On retrouve les mêmes caractéristiques pour les systèmes, avec toutefois comme facteur le plus marqué le nombre de sens du mot-cible : plus celui-ci est polysémique plus les scores ont tendance à baisser, cette caractéristique n'étant pas significative pour les annotateurs. On peut l'expliquer par la difficulté qu'ont les systèmes à proposer des substituts différents pour les différents sens des mots-cibles : bien souvent les réponses sont très proches et les substituts proposés sont identiques.

S'il est bien entendu possible d'envisager de nombreuses interactions entre les variables, nous avons avant tout voulu observer les trois sous-ensembles du jeu de test correspondant aux trois catégories grammaticales des mots-cibles. Les résultats sont indiqués dans la table 2.

On constate que les caractéristiques impliquées diffèrent d'une catégorie à l'autre. Si la difficulté de

Mesure	Adjectifs	Noms	Verbes
Entropie	rang du sens (0,42), nb. de synonymes (0,22)	fréq. mot-cible (0,41) , fréq. mots phrase (0,29), nb. de synonymes (0,23)	rang du sens (0,36), fréq. mots phrase (0,26)
OOT	rang du sens (0,45), nb. de synonymes (0,27)	nombre de sens (0,43), fréq. mots phrase (0,28), nb. de synonymes (0,22), mot-cible à droite (0,20), fréq. mot-cible (0,20)	nombre de sens (0,49), nb. de synonymes (0,29), rang du sens (0,27), complexité phrase (0,22)

TABLE 2 – Facteurs de la difficulté par catégorie (par r décroissant)

traitement des adjectifs est liée aux facteurs que nous venons d'évoquer, d'autres caractéristiques, cette fois liées à la phrase, émergent dans le cas des noms et des verbes. On observe qu'une phrase dont les mots sont très fréquents sera plus difficile à traiter par les humains et, dans le cas des noms, par les systèmes. Cela peut s'expliquer par le fait que les mots du contexte sont peu discriminants. Enfin, d'autres facteurs contextuels ont un impact sur les systèmes. La complexité syntaxique est assez logiquement un obstacle pour une analyse fine du contexte (notamment si les mots syntaxiquement reliés sont à grande distance de la cible, ce qu'une approche par fenêtre graphique aura du mal à capter) : il est donc logique d'observer un impact plus grand pour les verbes, même si la tendance générale va dans ce sens pour les autres catégories. La position du mot-cible a un impact pour le traitement du nom. On peut faire l'hypothèse que la cohésion lexicale dans la phrase est plus forte autour d'un nom en position topicale, ce qui faciliterait son traitement.

5 Conclusion

Ces résultats montrent le rôle central des caractéristiques inhérentes au mot-cible pour expliquer la dispersion des annotations. C'est le cas de la fréquence du mot, du nombre de synonymes, et surtout de la fréquence des sens d'un même mot-cible. Ce sont donc les mots polysémiques, fréquents et pour lesquels les synonymes disponibles sont nombreux qui posent le plus de problème dans ce type de tâche manuelle. Peu de caractéristiques liées à la phrase, et en tout cas aucune de celles qu'on a mesurées, ne semblent avoir d'importance centrale. Les performances des systèmes sont également liées aux caractéristiques sémantiques des mots, et en particulier, outre les traits que nous venons de citer, le degré de polysémie du mot-cible. Cependant, l'examen des performances des systèmes fait plus nettement apparaître des facteurs liés à la phrase. Notons que des corrélations peu marquées sont sans doute explicables par certains choix faits en amont pour la sélection du jeu de test, conçu de façon à sélectionner des phrases ayant de bonnes propriétés pour l'annotation (mots polysémiques ayant une fréquence minimale, phrases bien formées et autonomes).

De façon encore exploratoire, nous avons construit deux modèles linéaires multiples, pour prédire le OOT et l'entropie à partir des variables étudiées. Le modèle du score *OOT* est bien plus efficace ($R^2 = 0,37$, soit 37% de la variance expliquée) que celui de l'entropie (19%) : ceci montre bien que de nombreux autres facteurs interviennent, surtout pour l'annotation humaine. L'étude des résidus de ces modèles nous permet également d'identifier des pistes pour affiner l'identification des caractéristiques liées à cette tâche. C'est le cas notamment de la notion de figement ou de prototypicité de certains emplois : ainsi, *éplucher des pommes de terre* ou *interpréter le rôle* ont entraîné des accords supérieurs à ceux prévus par le modèle. A l'inverse, des collocations moins marquées ont eu l'effet inverse, comme *fonder la responsabilité*, *archétype grossier*, *direction idéologique*.

Références

B. BIGI, Ed. (2014). *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille. ATALA, LPL.

CARMEL D. & YOM-TOV E. (2010). *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan and Claypool.

FABRE C., HATHOUT N., HO-DAC L.-M., MORLANE-HONDÈRE F., MULLER P., SAJOUS F., TANGUY L. & VAN DE CRUYS T. (2014). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In (Bigi, 2014).

LAROUSSE (2005). *Le Grand Larousse Illustré - dictionnaire encyclopédique en 3 volumes*. Larousse.

MCCARTHY D. & NAVIGLI R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, **43**(2), 139–159.

E. MORIN & Y. ESTÈVE, Eds. (2013). *Actes de TALN 2013 (Traitement automatique des langues naturelles)*, Sables d'Olonne. ATALA, LINA.

PALMER M., DANG H. T. & FELLBAUM C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, **13**(2), 137–163.

PASSONNEAU R. J., SALLEB-AOUSSI A., BHARDWAJ V. & IDE N. (2010). Word sense annotation of polysemous words by multiple annotators. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, **39**(1), 161–182.

PRUVOST J. (2006). *Les dictionnaires français : outils d'une langue et d'une culture*. Ophrys.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In (Morin & Estève, 2013).

TANGUY L. & TULECHKI N. (2009). Sentence Complexity in French : a Corpus-Based Approach. In *Proceedings of of the International Conference on Recent Advances in Intelligent Information Systems (IIS)*, p. 131–145, Krakow, Poland.

URIELI A. & TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In (Morin & Estève, 2013).