Estimation de l'homogénéité sémantique pour les Questionnaires à Choix Multiples

Van-Minh Pho[1,2] Brigitte Grau[1,3] Anne-Laure Ligozat[1,3]
(1) LIMSI-CNRS, Orsay
(2) Université Paris-Sud, Orsay
(3) ENSIIE, Evry
prenom.nom@limsi.fr

Résumé. L'homogénéité sémantique stipule que des termes sont sémantiquement proches mais non similaires. Cette notion est au cœur de travaux relatifs à la génération automatique de questionnaires à choix multiples, et particulièrement à la sélection automatique de distracteurs. Dans cet article, nous présentons une méthode d'estimation de l'homogénéité sémantique dans un cadre de validation automatique de distracteurs. Cette méthode est fondée sur une combinaison de plusieurs critères de voisinage et de similarité sémantique entre termes, par apprentissage automatique. Nous montrerons que notre méthode permet d'obtenir une meilleure estimation de l'homogénéité sémantique que les méthodes proposées dans l'état de l'art.

Abstract.

Semantic homogeneity estimation for MCQs

Semantic homogeneity states that terms are semantically close but not similar. This notion is the focus of work related to multiple-choice test generation, and especially to automatic distractor selection. In this paper, we present a method to estimate semantic homogeneity within a framework of automatic distractor validation. This method is based on a combination of several criteria of semantic relatedness and similarity between terms, by machine learning. We will show that our method allows to obtain a better estimation of semantic homogeneity than methods proposed in related work.

Mots-clés : similarité, voisinage sémantique, classification de termes.

Keywords: similarity, semantic relatedness, term ranking.

1 Introduction

La notion de similarité sémantique permet d'estimer l'intensité du lien sémantique reliant deux concepts. Cette notion est au cœur de travaux relatifs à la génération automatique de Questionnaires à Choix Multiples (QCM), et particulièrement à la sélection automatique de distracteurs (mauvais choix de réponse) , qui se doivent d'être sémantiquement homogènes pour respecter les règles de construction d'un QCM. Ces travaux (Karamanis *et al.*, 2006; Mitkov *et al.*, 2009) abordent la problématique de la sélection automatique de distracteurs comme un problème de similarité sémantique entre les différentes options (choix de réponse incluant les distracteurs et la réponse correcte). Cependant, la problématique de la sélection automatique de distracteurs est différente d'un problème de similarité sémantique : en effet, bien que les options doivent être sémantiquement proches, elles ne doivent pas être de sens similaires.

Nous proposons donc une définition différente de la notion d'homogénéité sémantique qui stipule que les options sont sémantiquement proches mais non similaires. L'objectif de cet article est de proposer une méthode d'estimation de l'homogénéité sémantique des options en combinant différentes mesures de voisinage et de similarité sémantique, dans une perspective applicative de validation ou de génération automatique de QCM. Afin de présenter en détail la notion d'homogénéité sémantique et la méthode que nous proposons, il est nécessaire d'exposer le contexte des QCM.

Les QCM sont largement utilisés dans de nombreux contextes d'apprentissage et d'évaluation. Les principales raisons en sont que leur évaluation peut être automatisée et que leur pertinence, ainsi que leur objectivité dans l'évaluation des compétences de l'apprenant, ont été prouvées (Haladyna *et al.*, 2002). Cependant, la rédaction de QCM est coûteuse

en temps, et la qualité des QCM est cruciale si l'on veut s'assurer que les résultats des apprenants correspondent à leurs compétences. Ainsi, la réalisation d'applications permettant l'évaluation automatique de la qualité de QCM ou la génération automatique de ceux-ci est nécessaire pour aider les enseignants.

Un QCM est un ensemble de *questions*, chacune d'entre elles étant composée de deux parties (cf. exemple ci-dessous) : l'*amorce* et les *options*, incluant la *réponse* (option correcte) et un ou plusieurs *distracteurs* (options incorrectes).

Amorce : De quel pays est originaire le kimchi ?
Réponse : Corée
Distracteur : Japon
Distracteur : Chine
Distracteur : Mongolie

La sélection des distracteurs est une tâche difficile lors de la création de QCM : la qualité d'une question dépend principalement de la qualité de ses options (Rodriguez, 2005). Le critère principal de la qualité des options est l'homogénéité sémantique de celles-ci, c'est-à-dire que les options doivent partager différents traits sémantiques. Cependant, elles doivent avoir des contenus sémantiques suffisamment différents pour constituer des réponses plausibles mais non possibles. Cette notion d'homogénéité sémantique découle des règles de rédaction de QCM «*Rendre la formulation des options homogène en contenu et en structure grammaticale*» et «*Rendre les options indépendantes les unes des autres : le sens de l'une ne doit pas être inclus dans le sens de l'autre*» (Haladyna *et al.*, 2002). Aussi, étant donné des candidats, extraits de différentes ressources, l'estimation de leur homogénéité sémantique par rapport à la réponse correcte permet de les filtrer et de ne garder que ceux qui sont assez homogènes pour constituer des distracteurs pertinents.

Nous proposons d'évaluer l'homogénéité sémantique en comparant chaque distracteur à la réponse et en calculant plusieurs critères fondés sur l'utilisation de différentes ressources sémantiques pour couvrir de nombreuses relations sémantiques, mesures qui sont combinées dans un modèle d'apprentissage automatique. Nous étendons les travaux existants (Karamanis *et al.*, 2006; Lee & Seneff, 2007; Mitkov *et al.*, 2009) en proposant une plus large palette de mesures, étendant ainsi leur couverture et en traitant plus de types de distracteurs, tels que les chunks (syntagmes de plus bas niveau) et les entités nommées, sans limitation de domaine. Dans cet article, nous montrerons par une évaluation de corpus que notre combinaison de mesures de voisinage et de similarité sémantique permettent d'obtenir une meilleure estimation de l'homogénéité sémantique que les méthodes présentées par les travaux précédents.

2 État de l'art

Il existe différentes stratégies d'estimation de l'homogénéité sémantique : mesure fondée sur la fréquence des collocations (Lee & Seneff, 2007), mesures de voisinage ou de similarité distributionnel (Karamanis *et al.*, 2006; Mitkov *et al.*, 2009), sélection des distracteurs appartenant à un document de référence à partir duquel la question (amorce et réponse) est créée (Mitkov *et al.*, 2009), mesures fondées sur des bases de connaissances hiérarchiques (Mitkov *et al.*, 2009) ou sur la similarité phonétique (Mitkov *et al.*, 2009). La mesure calculée par (Lee & Seneff, 2007) privilégie les termes apparaissant le plus fréquemment avec les contextes gauche et/ou droit de la réponse extraite d'une phrase. (Karamanis *et al.*, 2006) et (Mitkov *et al.*, 2009) calculent un score de similarité distributionnelle entre les candidats et la réponse. (Mitkov *et al.*, 2009) calculent notamment une mesure dont les co-occurrents comparés sont les mots liées aux mots comparés par une relation de dépendance dans un grand corpus de documents. (Mitkov *et al.*, 2009) utilisent d'autres stratégies : la première consiste à privilégier les candidats apparaissant dans le document de référence. Les autres stratégies sont les mesures de voisinage et de similarité sémantique fondées sur WordNet exposées à la section 3.3 et une mesure de similarité phonétique fondée sur Soundex, un algorithme d'indexation phonétique de mots.

L'estimation de l'homogénéité sémantique est évaluée à travers la qualité des distracteurs sélectionnés. Cette évaluation est effectuée par des apprenants (à travers des tests psychométriques) (Lee & Seneff, 2007; Mitkov *et al.*, 2009) ou par le jugement d'experts du domaine (Karamanis *et al.*, 2006).

Les principales limitations de ces travaux sont qu'ils sont limités à une application de domaine spécifique (médecine, apprentissage des prépositions) et/ou par les types syntaxiques des réponses (mots, syntagmes nominaux limités aux chunks nominaux suivis ou non d'un chunk prépositionnel à tête nominale (Mitkov & Ha, 2003)) tandis que notre travail n'est pas spécifique à un domaine et qu'il couvre tout type de chunk et d'entité nommée. De plus, aucun des travaux de l'état de l'art n'évalue l'homogénéité sémantique sur un corpus de QCM de référence.

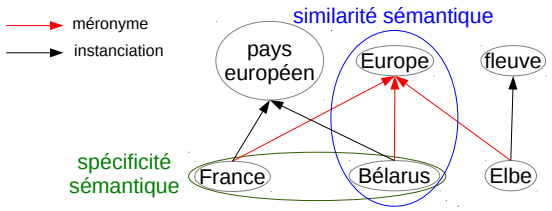

FIGURE 1 – Caractérisation sémantique de paires de nœuds

3 Homogénéité sémantique

L'homogénéité sémantique stipule que les options partagent des caractéristiques sémantiques communes. Dans l'exemple de la section 1, toutes les options sont des pays asiatiques. L'homogénéité sémantique des différentes options signifie que celles-ci doivent être sémantiquement voisines, sans partager le même contenu sémantique.

Nous donnons la définition de plusieurs notions utiles à la définition de l'homogénéité sémantique en faisant référence à une organisation des connaissances sous la forme d'un graphe hiérarchique (figure 1) contenant des concepts typés et des relations sémantiques.

La définition du *voisinage sémantique* est la suivante : «Le voisinage sémantique indique dans quelle mesure deux concepts sont sémantiquement distants dans un réseau ou une taxonomie en utilisant toutes les relations entre eux (c'est-à-dire des relations d'hyponymie/hyperonymie [1], d'antonymie [2], de méronymie [3] et toutes sortes de relations fonctionnelles incluant *is-made-of*, *is-an-attribute-of*, etc.)» (Ponzetto & Strube, 2007). Le voisinage sémantique est établi entre deux termes lorsqu'il existe un chemin entre les concepts auxquels ils se réfèrent, et le degré de voisinage est dépendant de la longueur du chemin. Dans la figure 1, tous les concepts peuvent être considérés comme sémantiquement voisins.

Nous définissons la *similarité sémantique* comme un cas particulier de voisinage sémantique : deux termes sont similaires s'ils partagent le même sens (c'est-à-dire qu'ils sont des synonymes) ou un sens partiel, c'est-à-dire que les concepts auxquels ils se réfèrent sont liés par une chaîne ascendante ou descendante de relations *is-a* ou de méronymie, tels que «Bélarus» et «Europe» dans la figure 1.

Nous définissons la *spécificité sémantique* comme un cas particulier de voisinage sémantique entre deux termes dont les concepts auxquels ils se réfèrent partagent un ancêtre commun direct, à l'instar de «France» et «Bélarus» dans la figure 1.

Nous proposons enfin la définition de l'*homogénéité sémantique* comme étant un cas particulier de *voisinage sémantique* qui considère toutes les relations entre les concepts comparés mais exclut la notion de *similarité sémantique* : deux options ne peuvent être similaires. Enfin, une meilleure homogénéité est atteinte si la *spécificité sémantique* est respectée.

Pour évaluer l'homogénéité sémantique des options, nous comparons les distracteurs à la réponse sur différents traits sémantiques. Notre objectif n'est pas d'apprendre une décision indiquant une homogénéité sémantique ou non car celle-ci est plus une question de degré qu'une décision binaire. De ce fait, nous considérons l'estimation automatique de l'homogénéité sémantique comme un problème d'ordonnancement dans lequel les candidats sémantiquement homogènes, soit les candidats classés dans les premiers rangs, sont les distracteurs. Les candidats à classer, excepté les distracteurs, sont sélectionnés selon des critères syntaxiques.

Pour estimer l'homogénéité sémantique entre un distracteur et la réponse, nous calculons plusieurs scores de voisinage et de similarité sémantique. Ces mesures sont calculées à partir de différents types de ressources : des représentations sémantiques hiérarchiques, et des représentations distributionnelles. Les représentations hiérarchiques permettent de prendre en compte les relations sémantiques explicites pour comparer deux concepts. Cependant, les ressources correspondant à ces représentations sont souvent limitées par leur couverture. Les mesures de voisinage distributionnel, construites selon le principe stipulant que des termes ayant des contextes similaires dans un corpus sont sémantiquement voisins, ont une couverture plus large mais la nature des relations sémantiques et les informations sur les représentations hiérarchiques sont inconnues. Dans les sections suivantes, nous présentons chacune des mesures choisies.

1. Deux concepts dont le premier a un sens plus spécifique/général que le second.
2. Deux concepts dont les sens sont opposés.
3. Deux concepts dont le premier est une partie ou un membre du second.

Les mesures fondées sur des représentations hiérarchiques ou des concepts explicitement typés sont :
— la similarité des types d'entité nommée ;
— la similarité des types spécifiques provenant de la base de connaissances DBpédia ;
— plusieurs mesures de voisinage sémantique fondées sur la base de données lexicale WordNet.

Les mesures de voisinage distributionnel sont :
— la comparaison des liens de Wikipédia ;
— l'Analyse Sémantique Explicite.

3.1 Similarité des types d'entité nommée

La première mesure de voisinage sémantique que nous présentons est fondée sur une annotation de texte en entités nommées, c'est-à-dire classiquement les noms de lieux, personnes et organisations. Nous considérons que l'identité de type d'entité nommée de deux termes est théoriquement un critère nécessaire pour qu'ils soient sémantiquement homogènes. Ce critère n'est pas suffisant car les types d'entité nommée choisis sont des catégories très générales, et qu'il convient également d'exclure les termes similaires.

Afin de mesurer ce critère d'homogénéité, nous considérons les trois grandes catégories : *lieu*, *organisation* et *personne*. Pour comparer le type d'entité nommée de deux termes, nous utilisons la mesure binaire $meme\_type\_EN$ (équation (1)) qui indique simplement si les termes sont de même type d'entité nommée ou non.

$$meme\_type\_EN(t_1, t_2) = \begin{cases} 1 \text{ si } EN(t_1) = EN(t_2) \wedge t_1 \text{ est une entité nommée} \wedge t_2 \text{ est une entité nommée} \\ 0 \text{ sinon} \end{cases} \quad (1)$$

où t_1 et t_2 sont deux termes et $EN(t)$ est le type d'entité nommée du terme t.

3.2 Similarité des types sémantiques provenant de DBpédia

En plus de mesurer la similarité de types d'entité nommée généraux, nous souhaitons comparer les types sémantiques des termes à un niveau de granularité plus fin : des types plus spécifiques permettent de vérifier plus précisément l'homogénéité sémantique. Ainsi dans l'exemple de la section 1, les options sont toutes des pays asiatiques, ce qui donne une indication plus précise de leur homogénéité sémantique que de vérifier simplement qu'elles sont des lieux.

Cependant, tandis que les types d'entité nommée peuvent être reconnus indépendemment d'une ressource, les types spécifiques doivent être reconnus à partir d'une taxonomie hiérarchique. Pour cela, nous utilisons DBpédia [4], une ressource hiérarchique construite à partir des pages de Wikipédia. Les entités de DBpédia sont associées à des types sémantiques qui représentent les classes de l'ontologie de DBpédia, organisées en une taxonomie [5]. Cette ressource a l'avantage d'avoir une large couverture pour le domaine ouvert.

Pour calculer l'homogénéité sémantique entre deux termes fondée sur leur type DBpédia et leur position dans leur taxonomie, nous utilisons la mesure de Wu et Palmer (Wu & Palmer, 1994), $wup(t_1, t_2)$.

$$wup(t_1, t_2) = \frac{2 \times profondeur(lcs(t_1, t_2))}{profondeur(type(t_1)) + profondeur(type(t_2))} \quad (2)$$

où $type(t)$ est le type DBpédia du terme t, $profondeur(u)$ est la profondeur du type u dans la taxonomie ($profondeur(u) = 1$ si u est la racine de la taxonomie) et $lcs(t_1, t_2)$ est l'hyperonyme commun le plus spécifique de $type(t_1)$ et $type(t_2)$. La mesure de Wu et Palmer est fondée sur le chemin le plus court entre les deux concepts pondéré par leur profondeur dans la taxonomie. Ainsi, deux concepts profonds de parent commun obtiennent un meilleur score que deux concepts moins profonds de parent commun.

4. http://dbpedia.org/About
5. http://mappings.dbpedia.org/server/ontology/classes/

3.3 Mesures de voisinage sémantique fondées sur WordNet

Les mesures précédentes sont fondées sur la similarité des types sémantiques des termes. Dans cette sous-section, nous présentons des mesures de voisinage et de similarité sémantique fondées sur le sens des termes et les relations qui les lient.

Pour mesurer l'homogénéité sémantique sur tout type de termes et particulièrement les termes qui ne sont pas des entités nommées, nous utilisons des mesures définies pour WordNet [6], un réseau lexical qui groupe les mots synonymes en *synsets* liés par des relations sémantiques. Une glose (définition) est associée à chaque synset. WordNet contient également des entités nommées, mais restreintes à quelques types d'entité nommée (grandes villes, pays, continents...). Nous utilisons les quatre mesures sélectionnées par (Mitkov *et al.*, 2009) dans leur travail de sélection automatique des distracteurs : la mesure de recoupement étendu de gloses (*extended gloss overlap measure*) (Banerjee & Pedersen, 2003) ; la mesure de Leacock et Chodorow (Leacock & Chodorow, 1998) fondée sur le plus court chemin entre les synsets ; les mesures de Jiang et Conrath (Jiang & Conrath, 1997), et de Lin (Lin, 1997), fondées sur le *contenu d'information* (*information content*).

La mesure de recoupement étendu de gloses (Banerjee & Pedersen, 2003) ($simREG$) prend en compte les gloses des synsets comparés, ainsi que les gloses de leurs hyperonymes et de leurs hyponymes.

$$\begin{aligned} simREG(s_1, s_2) = score(def(s_1), def(s_2)) + score(hype(s_1), hype(s_2)) + score(hypo(s_1), hypo(s_2)) \\ + score(hype(s_1), def(s_2)) + score(def(s_1), hype(s_2)) \end{aligned} \quad (3)$$

où $def(s)$ est la glose du synset s, $hype(s)$ est la glose de l'hyperonyme de s (si s a plusieurs hyperonymes, alors $hype(s)$ est la concaténation de ces gloses) et $hypo(s)$ est la glose de l'hyponyme de s (si s a plusieurs hyponymes, alors $hypo(s)$ est la concaténation de ces gloses). $score(g(s_1), g(s_2))$ est la somme des longueurs des chaînes communes des gloses de s_1 et s_2 au carré.

La mesure de Leacock et Chodorow (Leacock & Chodorow, 1998) ($simLCH$) est fondée sur le plus court chemin entre deux synsets dans la taxonomie, c'est-à-dire le nombre minimal d'arêtes entre ces synsets. Cette mesure ne prend en compte que les relations d'hyperonymie et d'hyponymie.

$$simLCH(s_1, s_2) = -log(\frac{len(s_1, s_2)}{2 \times MAX}) \quad (4)$$

où $len(s_1, s_2)$ est le nombre minimal d'arêtes entre les synsets s_1 et s_2, et MAX est la profondeur de la taxonomie.

Les mesures de Jiang et Conrath (Jiang & Conrath, 1997) et de Lin (Lin, 1997) sont fondées sur le contenu d'information des synsets et de leur hyperonyme commun ($lcs(s_1, s_2)$). Le contenu d'information $IC(s)$ représente l'importance de l'information relative à un synset s dans un contexte donné.

$$IC(s) = -log(p(s)) \quad (5)$$

où $p(s)$ est la probabilité d'apparition de s et de ses hyponymes dans un corpus de référence. Ainsi, pour deux synsets s et s' tels que s' est un hyponyme de s, $p(s) \geq p(s')$ et donc $IC(s) \leq IC(s')$.

Les formules de la mesure de Jiang et Conrath ($simJCN$) et de Lin ($simLin$) sont présentées ci-dessous.

$$simJCN(s_1, s_2) = \frac{1}{IC(s_1) + IC(s_2) - 2 \times IC(lcs(s_1, s_2))} \quad (6)$$

$$simLin(s_1, s_2) = \frac{2 \times IC(lcs(s_1, s_2))}{IC(s_1) + IC(s_2)} \quad (7)$$

où $lcs(s_1, s_2)$ est l'hyperonyme commun le plus spécifique des synsets s_1 et s_2. Ces mesures comparent le contenu d'information des synsets s_1 et s_2 avec celui de leur hyperonyme commun le plus spécifique. Ces mesures combinent les connaissances sémantiques de WordNet avec les distributions des concepts dans un grand corpus. Elles privilégient les

6. http://wordnet.princeton.edu/

concepts dont l'hyperonyme commun le plus spécifique est proche, et dont les contenus d'informations (calculés à partir des distributions des concepts dans le corpus) sont proches de leur hyperonyme commun le plus spécifique.

Les termes pouvant avoir plusieurs sens, nous les associons à plusieurs synsets. Ainsi, pour calculer le voisinage sémantique entre deux termes, nous calculons les mesures sur toutes les paires de synsets associées aux termes et nous gardons le score maximal.

Ces mesures se complètent car elles calculent le score de similarité ou de voisinage sémantique selon différents critères : tandis que la mesure de Leacock et Chodorow se fonde uniquement sur les relations sémantiques explicites entre les concepts, la mesure de recoupement étendu de gloses se fonde sur la proximité textuelle des gloses des concepts et les mesures de Jiang et Conrath et de Lin se fondent sur un corpus de textes pour comparer l'importance (représenté par la fréquence d'apparition) des concepts.

Afin d'avoir également des mesures s'appliquant à tout type de termes sans la limite de leur présence ou non dans une ressource sémantique hiérarchique, nous avons sélectionné des mesures pouvant être calculées sur de grands corpus. Cette famille de mesures ne prennent pas en compte les relations sémantiques explicites, et sont dédiées à l'estimation du voisinage sémantique à l'instar des mesures précédentes concernant la mesure du voisinage sémantique. L'hypothèse qui a été considérée pour la conception de ces mesures est que les termes ont des sens sémantiquement proches s'ils partagent un contexte similaire.

3.4 Comparaison des liens de pages de Wikipédia

Une représentation contextuelle possible d'un terme est l'ensemble des liens entrants et sortants associés à une page de Wikipédia. Nous considérons les pages dont le titre correspond au terme d'une option. Les liens entrants et sortants représentent les pages de Wikipédia associées au corps d'une page de Wikipédia. L'outil Wikipedia Miner (Milne & Witten, 2013) calcule un score appris sur ces liens à partir de dumps de Wikipédia. Ce score est une combinaison de huit attributs représentant quatre mesures calculées sur les liens entrants et sortants :

— l'union des liens des pages comparées ;
— l'intersection des liens des pages comparées ;
— la *normalized link distance*, adaptée de la *normalized Google distance* (Cilibrasi & Vitanyi, 2007), calculant la distance sémantique de deux pages p_1 et p_2 en comparant les pages de Wikipédia où apparaissent les liens associés à p_1 et p_2. Si p_1 et p_2 sont liées aux mêmes pages, cela signifie un fort voisinage sémantique entre p_1 et p_2, tandis que si p_1 et p_2 sont liées à des pages différentes, cela signifie un faible voisinage sémantique entre p_1 et p_2. La formule de la *normalized link distance* est la suivante : $nld(p_1, p_2) = \frac{log(max(|P_1|,|P_2|)) - log(|P_1 \cap P_2|))}{log(|W|) - log(min(|P_1|,|P_2|))}$ où P_1 et P_2 sont les ensembles de pages reliant respectivement p_1 et p_2, et W est l'ensemble de toutes les pages de Wikipédia ;
— la similarité vectorielle des liens (*link vector similarity*), inspirée de la mesure du TF-IDF mais appliquée aux liens des pages comparées vers les pages de Wikipédia, au lieu des mots traités par le TF-IDF. Les dimensions des vecteurs comparés sont calculées à partir du *lf* $\times$ *iaf* (*link frequency* $\times$ *inverse article frequency*). La fréquence des liens (*link frequency*) donne une estimation de l'importance du lien l_i dans une page p_j : $lf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ où $n_{i,j}$ est le nombre d'occurrences du lien l_i dans la page p_j. La fréquence inverse d'article (*inverse article frequency*) mesure l'importance générale d'un lien : $iaf_i = log(\frac{|W|}{|p:l_i \in p|})$ où W est l'ensemble des pages de Wikipédia. Le score de voisinage sémantique est l'angle des vecteurs des pages comparées.

3.5 Analyse Sémantique Explicite

Une autre représentation contextuelle des termes est leur distribution à travers les documents d'un corpus. Deux termes dont les distributions (c'est-à-dire les fréquences d'apparition) sont proches dans les mêmes documents ont une forte probabilité d'être sémantiquement voisins. Afin de comparer les distributions de deux termes, nous calculons une mesure fondée sur l'Analyse Sémantique Explicite (*Explicit Semantic Analysis*, ESA) (Gabrilovich & Markovitch, 2007). L'ESA est fondée sur une représentation vectorielle de textes (d'un mot à un document entier) dont les dimensions sont les poids du texte dans chaque document du corpus. Un mot est représenté par un vecteur de poids et un texte contenant plusieurs mots est représenté par le barycentre des vecteurs de poids représentant chaque mot du texte. Le poids d'un mot m dans un document d correspond au TF-IDF de m dans d. Le score de voisinage de deux textes est le cosinus des vecteurs

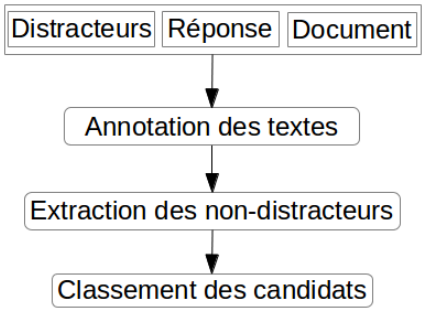

FIGURE 2 – Architecture montrant les différentes étapes de la validation automatique de distracteurs

représentant ces textes. Dans notre cas, le corpus de documents est Wikipédia. Pour calculer la mesure fondée sur l'ESA, nous utilisons l'outil ESAlib [7].

4 Évaluation de la qualité des distracteurs

Nous appliquons l'estimation de l'homogénéité sémantique des options à la validation automatique de distracteurs. Pour ce faire, nous joignons les distracteurs corrects de la question à des non-distracteurs (termes n'ayant pas été manuellement sélectionnés pour être les distracteurs de la question). Nous appelons *candidats* l'ensemble des distracteurs et des non-distracteurs. Notre objectif est d'apprendre un modèle d'ordonnancement capable de classer les distracteurs dans les premiers rangs, étant donné qu'ils devraient être plus homogènes à la réponse que les non-distracteurs. De plus ce modèle doit être capable de reconnaître les non-distracteurs similaires à la réponse mais non reconnus comme tels par la ressource WordNet afin de ne pas les sélectionner.

Les questions que nous traitons sont associées à un document de référence à partir duquel les amorces sont conçues. Les non-distracteurs sont sélectionnés dans ce document selon un critère d'homogénéité syntaxique. La figure 2 montre les différentes étapes de la validation automatique des distracteurs. Une première étape consiste à annoter le document afin d'obtenir des informations syntaxiques et sémantiques sur les options. Les non-distracteurs sont ensuite sélectionnés selon deux méthodes différentes (chacune de ces méthodes est évaluée séparément) : la première méthode consiste à extraire les non-distracteurs à partir des annotations du document (evalNDdocument), et la seconde consiste à sélectionner les options des autres questions du corpus (evalNDoptions). Les candidats obtenus sont classés afin de valider les distracteurs.

4.1 Annotation du document et des options

Pour extraire les non-distracteurs et calculer les différentes mesures, les candidats et les réponses doivent être annotés par des informations syntaxiques et sémantiques. L'annotation est de meilleure qualité si ces extraits sont analysés dans leur contexte. Ainsi, nous effectuons quatre annotations du document dans l'ordre suivant :

1. une analyse syntaxique avec l'outil Stanford Parser (Klein & Manning, 2003) ;
2. une annotation en entité nommées avec l'outil Stanford Named Entity Recognition (Finkel *et al.*, 2005) ;
3. une annotation en types spécifiques pour trouver les entités associées à des entités de DBpédia (et, par extension, des pages de Wikipédia). Cette annotation est effectuée avec l'annotateur DBpedia Spotlight (Daiber *et al.*, 2013), qui associe des entités de DBpédia aux termes correspondants du document, et désambiguïse ces termes si nécessaire. Cependant, certains termes ne sont pas annotés (à tort) par DBpedia Spotlight. Nous associons ces termes à toutes les entités de DBpédia dont les titres correspondent à ces termes, donc sans désambiguïsation ;
4. une annotation en synsets de WordNet, visant à associer les termes (candidats et réponse) avec un ou plusieurs synsets de WordNet, comme suit :

7. http://ticcky.github.io/esalib/

— si le terme apparaît dans WordNet, le terme est associé à ses synsets correspondants ;
— si le terme n'apparaît pas dans WordNet et n'est pas une entité nommée, elle est associée aux synsets correspondants à sa tête syntaxique (par exemple, le chunk «the little cat» est associé aux synsets de WordNet de nom «cat»).

Les annotations des options sont extraites de leurs correspondances dans le document. Si une option n'apparaît pas dans le document, elle est annotée hors contexte de la même manière que l'est le document.

4.2 Extraction et annotation des non-distracteurs

L'extraction des non-distracteurs est différente selon le type d'évaluation des distracteurs (evalNDdocument et evalNDoptions).

Dans le cas de l'évaluation evalNDdocument, les non-distracteurs sont extraits de la même source que la question (car les questions sont créées à partir d'un document de référence). Ces non-distracteurs sont tous syntaxiquement homogènes à la réponse. Si celle-ci est une entité nommée, les non-distracteurs sont les entités nommées du document annotées avec le Stanford Named Entity Recognition, étant donné que (Pho *et al.*, 2014) ont montré que les distracteurs ont généralement le même type d'entité nommée que la réponse. Néanmoins, afin de prendre en compte la métonymie, nous sélectionnons tous les non-distracteurs qui sont des entités nommées, quel que soit leur type d'entité nommée. Si la réponse est un chunk et n'est pas une entité nommée, les non-distracteurs sont les chunks du document de même type syntaxique que la réponse. Les chunks sont sélectionnés à partir des arbres de constituants des phrases du document, avec Tregex (Levy & Andrew, 2006), un outil permettant de sélectionner des nœuds d'arbres syntaxiques à partir de patrons. À ces non-distracteurs, nous associons leur type d'entité nommée, leurs entités de DBpédia et leurs synsets de WordNet annotés dans le document.

Dans le cas de l'évaluation evalNDoptions, les non-distracteurs sont les options des autres questions du corpus. Si la réponse à la question n'est pas une entité nommée, seuls les non-distracteurs de même type syntaxique que la réponse sont gardés.

Pour les deux évaluations, un dernier filtrage consiste à retirer les non-distracteurs similaires à une option, afin d'éviter les chevauchements sémantiques : deux éléments sont considérés comme similaires s'ils sont associés aux mêmes entités de DBpédia ou s'ils réfèrent aux mêmes synsets dans WordNet. Parmi les non-distracteurs sélectionnés, certains d'entre eux pourraient être assez pertinents pour être des distracteurs mais ne le sont pas car la question contient assez de distracteurs, ou sont des distracteurs d'autres questions. Dans cet article, nous traitons ces non-distracteurs comme des non-distracteurs normaux mais nous envisageons de faire une annotation manuelle des non-distracteurs afin d'écarter ces cas.

4.3 Ordonnancement sémantique

Le classement des candidats selon les différents critères d'homogénéité sémantique est effectué avec SVMRank [8], un outil d'ordonnancement automatique par apprentissage supervisé fondé sur un modèle SVM (*Séparateur à Vaste Marge* ou *Support Vector Machine*). Un SVM est un classifieur discriminant défini par un hyperplan séparant les données des différentes classes. L'outil SVMRank compare les couples de distracteurs-non-distracteurs d'une même question et apprend les poids des critères tels que pour chaque couple de distracteur-non-distracteur (d, nd), $svm(d) > svm(nd)$, où $svm(c)$ est le score attribué au candidat c à partir de la combinaison des critères et des poids de chacun de ces critères, appris par SVM. Pour l'évaluation evalNDoptions, nous ajoutons un critère supplémentaire indiquant si le candidat apparaît dans le document de référence de la question ou non.

5 Expériences

5.1 Corpus

Afin d'évaluer notre méthode, nous utilisons un corpus de QCM en langue anglaise extrait de différentes sources : des tests d'évaluation de systèmes de compréhension automatique de textes fournis par QA4MRE [9] (ensemble qa4mre) et

8. http://www.cs.cornell.edu/people/tj/svmlight/svmrank.html
9. http://www.celct.it/newsReader.php?idnews=74

corpus	ensemble	# q.	# opt.	(# q.)/opt.	objectif
tousQCM	qa4mre	341	1531	4,5	compréhension automatique de textes
	evalAnglais	394	1292	3,3	évaluation de la langue
	total	735	2823	3,8	
qcmEN	qa4mre	56	252	4,5	compréhension automatique de textes
	evalAnglais	47	150	3,2	évaluation de la langue
	total	103	402	3,9	
qcmNonEN	qa4mre	51	239	4,7	compréhension automatique de textes
	evalAnglais	100	342	3,8	évaluation de la langue
	total	151	581	3,8	

TABLE 1 – Caractéristiques des corpus : nom des corpus, nom des ensembles, nombre de questions, nombre d'options, nombre moyen d'options par question et objectif

plusieurs sites web d'apprentissage de la langue anglaise (ensemble evalAnglais). L'ensemble qa4mre a été conçu pour évaluer des machines, mais (Pho *et al.*, 2014) montrent qu'ils respectent des critères de formation de QCM. À partir de ce corpus, nous avons établi deux sous-corpus : le premier est constitué de questions dont les réponses sont des entités nommées (corpus qcmEN présenté au tableau 1), à l'instar de la question suivante :

Énoncé : Which Japanese city was the first to try limit convenience store hours ?
Réponse : Kyoto
Distracteur : Saitama
Distracteur : Tokyo

et le second est constitué de questions dont les réponses sont des chunks qui ne sont pas des entités nommées (corpus qcmNonEN présenté au tableau 1), à l'instar de la question suivante :

Énoncé : Trade union officials fear that this new campaign might end up unjustly penalizing _____, by driving the employers further underground.
Réponse : workers
Distracteur : employers
Distracteur : the "Mr. Bigs"

Le tableau 1 montre plusieurs caractéristiques du corpus (tousQCM), ainsi que des deux sous-corpus sur lesquels nous travaillons : qcmEN et qcmNonEN.

Le corpus qcmEN comporte 14 % des questions du corpus tousQCM et le corpus qcmNonEN comporte environ 20 % du corpus tousQCM. Les questions que nous traitons (chunks et EN) composent plus d'un tiers du corpus d'origine, ce qui montre que ces types de questions sont couramment posées lors de tests. Sur chacun des sous-corpus qcmEN et qcmNonEN, l'apprentissage du modèle a été effectué séparément.

5.2 Méthode d'évaluation

Nous considérons que les distracteurs sont sémantiquement plus proches de la réponse que les non-distracteurs et, par conséquent, devraient avoir un meilleur rang. Afin d'évaluer cela, nous calculons la précision (équation (8)) et le rappel (équation (9)) moyens au rang n en fonction du nombre de distracteurs, ainsi que la f-mesure (équation (10)).

$$PM = \frac{\sum_i^{nbQ} P_{i,nbD}}{nbQ} \quad (8) \qquad RM = \frac{\sum_i^{nbQ} R_{i,nbD}}{nbQ} \quad (9) \qquad F = 2 \times \frac{PM \times PR}{PM + PR} \quad (10)$$

où nbQ est le nombre de questions du corpus, nbD le nombre de distracteurs de la question évaluée, et $P_{i,nbD}$ et $R_{i,nbD}$ sont la précision (équation (11)) et le rappel (équation (12)) de la question i.

$$P_{i,nbD} = \frac{\#\text{D de rang} \leq nbD}{\#\text{C de rang} \leq nbD} \quad (11) \qquad R_{i,nbD} = \frac{\#\text{D de rang} \leq nbD}{nbD} \quad (12)$$

où D signifie distracteurs et C signifie candidats.

	qcmEN			qcmNonEN		
	R	P	F	R	P	F
$meme\_type\_EN$	0,83	0,26	0,40			
wup sur les types de DBpédia	0,70	0,34	0,46	0,94	0,14	0,24
$simREG$	0,67	0,25	0,36	0,37	0,23	0,28
$simLCH$	0,73	0,27	0,39	0,42	0,22	0,29
$simJCN$	0,83	0,23	0,36	0,40	0,18	0,25
$simLin$	0,84	0,23	0,36	0,40	0,18	0,25
Comparaison des liens de Wikipédia	0,41	0,32	0,36	0,76	0,22	0,34
ESA	0,40	0,34	0,37	0,35	0,24	0,28
Modèle d'ordonnancement	0,48	**0,46**	**0,47**	0,39	**0,36**	**0,37**

TABLE 2 – Résultats des méthodes de voisinage sémantique avec l'évaluation evalNDdocument

Cette évaluation constitue une proposition originale, les travaux étant usuellement évalués par des utilisateurs, ne permettant pas leur reproductibilité.

La précision et le rappel sont calculés pour chacune des mesures de voisinage sémantique, ainsi que pour le modèle d'ordonnancement. Nous évaluons le classement par une validation croisée en 7 sous-ensembles, c'est-à-dire que chacun des sous-ensembles du corpus est évalué selon le modèle appris à partir des autres sous-ensembles du corpus.

5.3 Résultats

Dans le cas de l'évaluation evalNDdocument, le tableau 2 montre que le modèle d'ordonnancement obtient un meilleur équilibre entre le rappel et la précision que les mesures individuelles, quel que soit le corpus. Le modèle donne une meilleure précision que les autres mesures et de meilleurs résultats que les mesures fondées sur WordNet, utilisées par (Mitkov *et al.*, 2009).

Certaines mesures évaluées donnent un meilleur rappel que le modèle d'ordonnancement. Nous distinguons deux cas : le premier concerne les mesures fondées sur les types (d'entité nommée et spécifiques) qui sont plus efficaces pour filtrer les candidats que pour sélectionner les distracteurs. Le second cas concerne les mesures dont la couverture des ressources est faible (WordNet dans le corpus qcmEN et Wikipédia dans le corpus qcmNonEN).

Les mesures donnent globalement des résultats inférieurs dans le corpus qcmNonEN. La raison principale est que les candidats et les réponses qui ne sont pas des entités nommées sont associés à moins d'informations sémantiques que les entités nommées, particulièrement sur les types sémantiques.

Dans le corpus qcmEN, la plupart des cas où les non-distracteurs ont un meilleur rang que les distracteurs sont dus au fait que les distracteurs et la réponse ne sont pas typés par un type (de DBpédia) très spécifique. Parmi les non-distracteurs restants, ceux-ci sont assez pertinents pour être des distracteurs ou sont similaires à la réponse, donc ne peuvent être des distracteurs.

La majorité des non-distracteurs du corpus qcmNonEN de meilleur rang que les distracteurs sont clairement des non-distracteurs mais certaines mesures (particulièrement celles fondées sur WordNet) considèrent que ces non-distracteurs sont sémantiquement plus voisins que les distracteurs. Parmi les non-distracteurs restants, certains d'entre eux ne sont pas sémantiquement proche de la réponse dans le contexte courant (document de référence) ou sont assez pertinents pour remplacer les distracteurs.

Dans le cas de l'évaluation evalNDoptions, le tableau 3 montre que les mesures individuelles donnent une précision plus faible que pour l'évaluation evalNDdocument. Cela est dû à deux causes principales. Premièrement, cette évaluation extrait plus de non-distracteurs que l'évaluation evalNDdocument, donc les mesures de faible couverture et/ou fondées sur les types donnent un très fort rappel et une très faible précision. Deuxièmement, un grand nombre de non-distracteurs sont sémantiquement plus proches de la réponse que les distracteurs, mais n'ont pas été sélectionnés manuellement car ils n'apparaissent pas dans le contexte de la question, soit le document de référence. En revanche, quel que soit l'évaluation, le modèle d'ordonnancement donne les mêmes résultats pour le corpus qcmEN, contrairement au corpus qcmNonEN où l'évaluation evalNDdocument donne de meilleurs résultats.

Les résultats montrent que l'appartenance au document de référence est un critère important pour ordonner les entités

	qcmEN			qcmNonEN		
	R	P	F	R	P	F
$meme\_type\_EN$	0,83	0,03	0,05			
wup sur les types de DBpédia	0,51	0,07	0,13	0,89	0,04	0,07
$simREG$	0,58	0,10	0,17	0,32	0,15	0,20
$simLCH$	0,65	0,08	0,15	0,36	0,13	0,19
$simJCN$	0,73	0,05	0,09	0,37	0,12	0,18
$simLin$	0,72	0,05	0,10	0,36	0,11	0,17
Comparaison des liens de Wikipédia	0,34	0,24	0,28	0,67	0,15	0,28
ESA	0,27	0,21	0,23	0,30	0,18	0,22
Modèle d'ordonnancement	0,43	**0,42**	**0,42**	0,24	**0,22**	0,22

TABLE 3 – Résultats des méthodes de voisinage sémantique avec l'évaluation evalNDoptions

nommées, contrairement aux chunks non entité nommée. En effet, un grand nombre de candidats sont des mots ou des n-grammes «communs» qui se retrouvent dans plusieurs documents comme le mot «track», ce qui fait que le critère d'appartenance à un document n'améliore pas le modèle d'apprentissage au niveau des chunks non entité nommée.

6 Conclusion

Dans cet article, nous avons proposé une méthode d'estimation de l'homogénéité sémantique fondée sur la combinaison par apprentissage de plusieurs mesures de voisinage et de similarité sémantique. Dans le cadre d'application à la validation automatique de QCM, nous obtenons des résultats supérieurs aux méthode de l'état de l'art. Les mesures fondées sur la similarité du type des options permettent de donner une indication sur la similarité de leurs catégories sémantiques et les mesures fondées sur les relations sémantiques entre les termes ainsi que les mesures de voisinage distributionnel permettent d'affiner la reconnaissance de l'homogénéité sémantique. Dans des travaux futurs, nous souhaitons adapter notre approche à tout type de réponse.

7 Remerciements

Ce travail a été financé par Digiteo dans le cadre du projet Aneth.

Références

BANERJEE S. & PEDERSEN T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 3, p. 805–810.

CILIBRASI R. L. & VITANYI P. M. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383.

DAIBER J., JAKOB M., HOKAMP C. & MENDES P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, p. 121–124 : ACM.

FINKEL J. R., GRENAGER T. & MANNING C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 363–370 : Association for Computational Linguistics.

GABRILOVICH E. & MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, p. 1606–1611.

HALADYNA T. M., DOWNING S. M. & RODRIGUEZ M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, **15**(3), 309–333.

JIANG J. J. & CONRATH D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

KARAMANIS N., HA L. A. & MITKOV R. (2006). Generating multiple-choice test items from medical text : A pilot study. In *Proceedings of the fourth international natural language generation conference*, p. 111–113 : Association for Computational Linguistics.

KLEIN D. & MANNING C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, p. 423–430 : Association for Computational Linguistics.

LEACOCK C. & CHODOROW M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, **49**(2), 265–283.

LEE J. & SENEFF S. (2007). Automatic generation of cloze items for prepositions. In *INTERSPEECH*, p. 2173–2176.

LEVY R. & ANDREW G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, p. 2231–2234 : Citeseer.

LIN D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, p. 64–71 : Association for Computational Linguistics.

MILNE D. & WITTEN I. H. (2013). An open-source toolkit for mining wikipedia. *Artificial Intelligence*, **194**, 222–239.

MITKOV R. & HA L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, p. 17–22 : Association for Computational Linguistics.

MITKOV R., HA L. A., VARGA A. & RELLO L. (2009). Semantic similarity of distractors in multiple-choice tests : extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, p. 49–56 : Association for Computational Linguistics.

PHO V.-M., ANDRÉ T., LIGOZAT A.-L., GRAU B., ILLOUZ G. & FRANÇOIS T. (2014). Multiple choice question corpus analysis for distractor characterization.

PONZETTO S. P. & STRUBE M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, **30**, 181–212.

RODRIGUEZ M. C. (2005). Three options are optimal for multiple-choice items : A meta-analysis of 80 years of research. *Educational Measurement : Issues and Practice*, **24**(2), 3–13.

WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, p. 133–138 : Association for Computational Linguistics.