De la quenelle culinaire à la quenelle politique : identification de changements sémantiques à l'aide des Topic Models

Ingrid Falk, Delphine Bernhard, Christophe Gérard

De la quenelle culinaire à la quenelle politique : identification de changements sémantiques à l'aide des Topic Models

Ingrid Falk Delphine Bernhard Christophe Gérard
LiLPa, Université de Strasbourg
ifalk, dbernhard, christophegerard@unistra.fr

Résumé. Dans cet article nous employons le « topic modeling » pour explorer des chemins vers la détection automatique de l'apparition de nouveaux sens pour des mots connus. Nous appliquons les méthodes introduites dans (Lau *et al.*, 2012, 2014) à un cas de néologie sémantique récent, l'apparition du nouveau sens de *geste* pour le mot « quenelle ». Nos expériences mettent en évidence le potentiel de cette approche pour l'apprentissage des sens du mot, l'alignement des topics à des sens de dictionnaire et enfin la détection de nouveaux sens.

Abstract. In this study we explore topic modeling for the automatic detection of new senses of known words. We apply methods developed in previous work for English (Lau *et al.*, 2012, 2014) on a recent case of new word sense induction in French, namely the appearence of the new meaning of *gesture* for the word « quenelle ». Our experiments illustrate the potential of this approach at learning word senses, aligning the topics with dictionary senses and finally at detecting the new senses.

Mots-clés : topic models, induction de sens, néologie sémantique.

Keywords: topic models, word sense induction, semantic neologism.

1 Introduction

Dans cet article nous nous proposons de contribuer à la détection automatique de la néologie sémantique qui, traditionnellement, est définie comme un changement de sens sans changement de forme. Dans ce but, nous appliquons les méthodes basées sur les Topic Models, présentées dans (Lau *et al.*, 2012, 2014), à l'identification automatique du nouveau sens du signifiant « quenelle ». Ces méthodes sont très récentes et ont l'avantage d'être relativement simples à mettre en œuvre en comparaison à la majorité des travaux proposés jusqu'alors pour la détection automatique de la néologie sémantique. Toutefois, elles n'ont pour l'heure pas été appliquées au français.

Le signifiant « quenelle », désignant originellement une préparation culinaire, a acquis récemment un sens politique qui a suscité de nombreux écrits sur la Toile. Nous disposons ainsi d'un matériau textuel comprenant d'une part les usages classiques de ce mot (domaine culinaire) et d'autre part de ses usages très récents (domaine politique).

Cette étude ne porte donc que sur un type particulier de néologie sémantique bien distinct d'autres types existants. Le cas du mot « quenelle » ne relève ainsi ni de l'extension de sens (*arriver*), ni de la restriction de sens (*pondre*), ni de la conversion (*centenaire*, adj $>$ *centenaire*, nom), ni de l'antonomase (un *bordeaux*). Bien qu'il soit possible d'argumenter dans le sens d'une création homonymique, il semble que la symbolique liée au geste raciste de la quenelle motive d'en parler en termes de création métaphorique, par exemple sur le modèle de « caviar » (domaine culinaire $>$ domaine sportif).

2 Travaux liés

L'identification de la néologie sémantique reste problématique, et les méthodes proposées sont donc essentiellement semi-automatiques ou limitées à l'analyse manuelle d'exemples précis à l'aide de mesures statistiques. La méthode proposée par Reutenauer (2012) procède par annotation du corpus en sémèmes (lemmes de mots pleins issus des définitions lexicographiques fournies par le Trésor de la Langue Française Informatisé pour chaque mot plein du corpus). Les cooccurrents

les plus significatifs d'un mot étudié sont ensuite extraits par calcul de la spécificité (Reutenauer *et al.*, 2010). Si, en l'occurrence, le couplage de la sémantique différentielle et de la linguistique informatique permet bien de rendre compte d'une évolution de sens (extension / restriction), la sélection du candidat à la néologie reste manuelle. Les travaux de Boussidan & Ploux (2011) reposent sur le modèle ACOM (Ji *et al.*, 2003) qui utilise l'analyse factorielle pour fournir des représentations géométriques de la cooccurrence des mots dans un corpus. L'étude des patrons de cooccurrence permet de détecter des changements de sens. L'analyse des mots cooccurrents pour la détection de la néologie sémantique est également proposée par Cabré & Nazar (2011).

Dans notre cas précis le changement de domaine du mot « quenelle » se traduit par un changement de thème. Les méthodes très récentes basées sur les Topic Models permettent de modéliser de manière extrêmement simple le contexte thématique et son évolution. Elles paraissent ainsi particulièrement adaptées à notre cas d'étude, même si pour l'heure elles n'ont à notre connaissance été appliquées qu'à l'anglais.

Un « topic model » est un modèle probabiliste permettant de déterminer des sujets ou thèmes abstraits dans une collection de documents. Dans un tel modèle un document est considéré comme composé d'un ou plusieurs « topics » qui à leur tour sont vus comme distributions de probabilité sur l'espace des mots. Un « topic model » est génératif dans la mesure où il spécifie une procédure probabiliste pour la génération de documents. Pour créer un nouveau document on choisit d'abord une distribution de topics. Ensuite on choisit aléatoirement pour chaque mot de ce document un topic suivant cette distribution, et finalement un mot du topic sélectionné. Des techniques statistiques standard permettent d'inverser ce processus et d'inférer l'ensemble des topics ayant mené, selon ce paradigme, à la collection de documents observée (Steyvers & Griffiths, 2007).

Le modèle utilisé le plus fréquemment en TAL est LDA – Latent Dirichlet Allocation (Blei *et al.*, 2003), qui présente l'inconvénient que l'utilisateur doit fixer le nombre de topics à priori. Pour les expériences que nous présentons ici nous appliquons un modèle non-paramétré appelé HDP – Hierarchical Dirichlet Process (Teh *et al.*, 2006), où le nombre de topics est inféré en même temps que les topics.

Lau *et al.* (2012, 2014) appliquent le HDP à une tâche de détection des sens d'un mot. Leur approche peut se résumer comme suit. Dans un premier temps on constitue pour le mot observé un corpus composé de phrases contenant ce mot, phrases qui représentent les usages du mot étudié. On applique ensuite le topic modeling et on compare les topics inférés aux sens attestés dans des dictionnaires. Dans un deuxième temps ce corpus est divisé en un corpus de référence, représentant les usages connus, et un corpus dit nouveau supposé de contenir les usages avec un nouveau sens. On est ainsi en mesure d'étudier et comparer les topics (sens) avec lesquels le mot est employé dans les deux corpus.

3 Ressources

Corpus. Le corpus à la base de nos expériences se constitue d'une part des 164 paragraphes du corpus Le Monde (1987-2006) contenant le mot « quenelle ». Cette partie du corpus (*REF* pour la suite) représente l'usage classique, dans le domaine culinaire, du mot « quenelle ». D'autre part nous avons extrait 342 paragraphes contenant le mot « quenelle » des journaux suivants, disponibles en ligne : La Croix, Le Figaro, Le Monde, L'Équipe, Les Echos et Libération. Ces paragraphes proviennent de 160 articles datés entre septembre 2013 et janvier 2014.

Prétraitement. Pour l'apprentissage des Topic Models nous ne prenons en compte que les noms communs, verbes, adjectifs et adverbes présents dans Morphalou (Romary *et al.*, 2004). Si dans Morphalou une forme n'a qu'un seul lemme, elle est remplacée par celui-ci. La taille du vocabulaire obtenu de cette manière est de 2 430 mots et la taille moyenne des paragraphes collectés passe de 77.35 mots avant prétraitement à 25.91 mots après.

Dictionnaires. Dans la plupart des dictionnaires en ligne (le TLFi, Larousse et Reverso) seul le sens culinaire de « quenelle » est attesté. Voici par exemple la définition du TLFi :

> Préparation, en forme de boulette ou de petit cylindre, à base de viande, d'abats, ou de poisson finement hachés, incorporée à une pâte de farine ou de mie de pain, que l'on sert telle quelle accompagnée d'une sauce ou dans une garniture de pâté chaud : vol-au-vent, bouchée à la reine.

Seuls Wiktionary et Wikipedia attestent aussi le nouveau sens de *geste*. Les articles de Wikipédia nous semblant plus exhaustifs, c'est ceux-ci que nous utilisons dans nos expériences.

Hypothèses de travail. Puisque le sens *geste* n'est apparu qu'en 2005 [1] nous considérons que le corpus *REF* ne contient pas d'usages de « quenelle » dans le sens *geste*. Par contre, le corpus *NOUV* peut contenir à priori des usages dans les deux sens, *culinaire* et *geste*. Cependant, les fréquences d'usage (164 usages sur 10 ans pour le corpus *REF* contre 342 usages sur quelques mois pour le corpus *NOUV*) laissent supposer une prépondérance de l'usage *geste*.

4 Expériences et discussion des résultats

Du corpus regroupant *REF* et *NOUV* et en appliquant un « Hierarchichal Dirichlet Process » (HDP) nous avons inféré un « topic model » [2]. Rappelons (Section 2) que ce type de modèle apprend le nombre de topics en même temps que la distribution des mots en topics.

Les 6 topics résultants sont présentés dans le Tableau 1. Ces extraits sont déjà suffisants pour constater que les topics 3 et 5 sont clairement associés au sens *culinaire* de « quenelle » alors que les topics 1, 2 et 4 relèvent plutôt du sens *geste*. Par contre le topic 6 ne présente pas une interprétation aussi évidente.

T1	quenelle geste antisémite salut photo nazi bras humoriste français effectuer ...
T2	quenelle jean marier marque fion fond glisser petite sionisme déposer ...
T3	quenelle carte menu plat brochet cuisine rue compter veau dessert ...
T4	quenelle spectacle public geste jeune interdiction antisémite humoriste monde ordre ...
T5	quenelle brochet pain épices grand sauce chair cuisine écrevisse beurre ...
T6	vide quenelle couverts charges cru debout derrière docteur effet émission ...

TABLE 1: Les 6 topics appris sur le corpus *REF+NOUV*. Nous affichons les 10 mots les plus probables de la distribution.

En s'appuyant sur ces topics, le topic modeling permet d'analyser les textes du corpus et de déterminer la présence et la proportion des topics présents dans chaque paragraphe. Un exemple est présenté en Tableau 2.

Paragraphe Du côté associatif, la Ligue des droits de l'Homme (LDH) a accueilli la circulaire avec prudence, redoutant les risques d'une interdiction préalable qui pourrait « fédérer autour de Dieudonné une sympathie réactionnelle de ceux qui se considèrent opprimés ». En revanche, la LDH encourage à poursuivre en justice chaque atteinte à la loi de Dieudonné, déjà condamné à de multiples reprises pour antisémitisme ou injure raciale. SOS Racisme et l'Union des étudiants juifs de France (UEJF) ont emboîté le pas, annonçant qu'ils poursuivraient désormais toutes les "quenelles" effectuées dans des lieux où elles "ne laissent pas de doute" sur leur caractère antisémite. Le Conseil représentatif des institutions juives de France (Crif) a quant à lui appelé à la « mobilisation » dans la vingtaine de villes où l'humoriste doit se produire.
Après prétraitement : côté associatif ligue droit homme accueillir circulaire redouter risques interdiction fédérer autour sympathie considérer revanche encourager poursuivre justice atteinte loi condamné multiple reprises antisémitisme injure racial racisme union étudiant juif annoncer poursuivre quenelle effectuer lieu laisser doute caractère antisémite conseil représentatif institution juives appelé mobilisation ville humoriste produire
Distribution des topics : T1 56.00%, T2 44.00%

TABLE 2: Exemple de distribution de topics dans un paragraphe. Interprétation : La génération de ce texte s'explique selon le modèle par un tirage de mots du topic T1 et du topic T2 avec des pourcentages respectifs de 56% et 44%.

Dans ce qui suit nous allons d'abord analyser si ces topics correspondent aux usages de « quenelle » dans notre corpus (Section 4.1) avant d'explorer d'autres méthodes et mesures introduites dans (Lau *et al.*, 2012, 2014), plus spécifiquement : l'alignement sens ↔ topic (Section 4.2), la détermination du sens prédominant (Section 4.3) et enfin la détection de sens nouveaux (Section 4.4).

1. Wikipédia, http ://fr.wikipedia.org/wiki/Dieudonné#Le_geste_de_la_.C2.AB_quenelle_.C2.BB_et_autres_signes_de_ralliement, 30/04/2014
2. Pour toutes nos expériences nous nous sommes fortement appuyés sur les logiciels disponibles librement à `https://github.com/jhlau/predom_sense` et `https://github.com/jhlau/hdp-wsi`

4.1 Correspondances topics ↔ usages dans corpus

Le tableau 3 montre pour chaque topic, le nombre de paragraphes pour lesquels ce topic est prédominant. Est considéré prédominant dans un paragraphe le topic avec la plus grande probabilité dans la distribution de topics pour ce paragraphe (ainsi dans le tableau 2 le topic prédominant est le topic T1).

Corpus	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
REF	0	6	**118**	0	**36**	**4**
NOUV	**229**	**32**	9	**42**	29	1
Total	229	38	127	42	65	5

TABLE 3: Topics prédominants par paragraphes.

Nous voyons que la plupart (154 sur 164) des paragraphes du corpus *REF* ont comme topic prédominant les topics T3 et T5, ce qui correspond bien au sens *culinaire* de « quenelle ». D'autre part les topics prédominants de 303 des 342 paragraphes du corpus *NOUV* sont les topics T1, T2 et T4 qui relèvent du sens *geste*.

Les erreurs potentielles d'analyse thématique sont les paragraphes du corpus *NOUV* à topic prédominant *culinaire* et ceux du corpus *REF* à topic prédominant *geste*. Le corpus *REF* n'ayant pas de paragraphes à topic prédominant *geste*, nous nous tournons vers les 38 paragraphes à topic *culinaire* dans le corpus *NOUV*. 7 des 9 paragraphes à topic prédominant T3 suggèrent effectivement un sens *culinaire*. Les 2 paragraphes restants sont trop courts (2 respectivement 3 mots) pour permettre une évaluation du sens [3].

Les 5 des 29 paragraphes à sens prédominant T5, ne correspondant pas à un sens *culinaire* sont :

quenelle continue inquiéter
doublé quenelle
quenelle laisser sale goût
quenelle laisser sale goût
mauvaise quenelle

Si la brièveté de ces textes ou le contexte avec des connotations (ambiguës) *culinaires* (*goût*, *mauvaise* [4]) pourrait éventuellement justifier l'analyse erronée, il faut noter que ces cas sont néanmoins faciles à départager pour un juge humain. Notons aussi que seulement 5 paragraphes ont comme topic prédominant le topic 6, qui était particulièrement difficile à interpréter. Pour conclure nous constatons que globalement l'analyse par les topic models permet de bien départager les deux sens de « quenelle » dans notre corpus.

4.2 Correspondances sens ↔ topic

Dans cette section nous appliquons les méthodes décrites dans (Lau *et al.*, 2014) pour déterminer le sens prédominant d'une collection de documents. Dans ces expériences le sens est représenté par les entrées d'un dictionnaire, en l'occurrence les gloses de Wikipédia. Ces gloses sont converties en distributions (par l'estimation du maximum de vraisemblance) et mises en relation avec les topics. Nous calculons la similarité d'un sens de dictionnaire s_i et d'un topic t_j suivant l'équation suivante (Lau *et al.*, 2014) :

$$sim(s_i, t_j) = 1 - JS(S\|T) = 1 - \frac{1}{2}KL(S\|M) - \frac{1}{2}KL(T\|M) \quad (1)$$

ou S et T sont respectivement les distributions multinomiales de mots pour les sens s_i et topics t_j respectivement ; $M = \frac{1}{2}(S+T)$; et $JS(X\|Y)$ et $KL(X\|Y)$ sont les divergences Jensen-Shannon et Kulbach-Leibler pour les distributions X et Y respectivement.

Nous présentons dans le Tableau 4 la similarité des topics, suivant l'équation (1), aux sens *culinaire* et *geste* de « quenelle » (tel qu'attesté dans Wikipédia). On voit à nouveau que cet alignement automatique correspond très bien à nos observations en Section 4.1. Il se pose cependant encore le problème de détecter automatiquement à partir de quel taux de similarité

3. Il s'agit ici de titres d'articles qui sont déjà courts à l'origine et ont été raccourcis davantage lors du prétraitement.
4. Pour les trois derniers textes il s'agit en fait de jeux de mots.

culinaire	T5 0.266	T3 0.2228	T1 0.1162	T2 0.1141	T6 0.0714	T4 0.0709	moyenne 0.1436
geste	T1 0.3917	T2 0.3181	T4 0.2559	T5 0.1722	T3 0.1289	T6 0.071	moyenne 0.223

TABLE 4: Topics les plus similaires aux sens *culinaire* et *geste*, tel qu'attesté dans Wikipédia et suivant équation (1)

un topic ne devrait plus être associé à un sens. Une possibilité serait éventuellement de prendre les valeurs au dessus de la moyenne.

4.3 Sens prédominant

Sur la base de cette similarité et de la distribution des topics dans les paragraphes nous déduisons une distribution des sens (suivant toujours (Lau *et al.*, 2014)) :

$$prevalence(s_i) = \sum_{j=1}^{T} \left(sim(s_i, t_j) \times \frac{f(t_j)}{\sum_{k=1}^{T} f(t_k)} \right) \tag{2}$$

ou T est le nombre de topics et $f(t_j)$ le nombre d'usages (paragraphes) à topic prédominant t_j. Suivant cette formule nous obtenons la prévalence des sens dans le corpus. Nous avons aussi calculé la proportion de paragraphes pour chaque sens selon notre analyse manuelle des données [5] :

sens	**prévalence**	**proportion paragraphes**
culinaire	0.3625	195/506 = 0.3854
geste	0.6375	≈304/506 = 0.6008

Le modèle permet donc d'estimer assez bien la proportion d'usages dans le corpus correspondant à un certain sens (de dictionnaire).

4.4 Détection de topics

Dans cette section nous appliquons une mesure de nouveauté introduite par Lau *et al.* (2012). Ce score permet d'estimer si certains topics sont inédits dans le nouveau corpus. Pour un topic t_i le score de nouveauté Nouv(t_i) (par rapport au mot « quenelle ») est définit par l'équation (3).

$$\text{Nouv}(t_i) = \frac{p_{NOUV}(t_i) - p_{REF}(t_i)}{p_{REF}(t_i)} \tag{3}$$

où $p_{NOUV}(t_i)$ et $p_{REF}(t_i)$ représentent les probabilités (calculées par estimation du maximum de vraisemblance) des topics t_i dans le corpus nouveau et de référence respectivement. Le score de nouveauté est élevé si un topic est beaucoup plus fréquent dans le nouveau corpus (que dans le corpus de référence) et est en même temps relativement peu fréquent dans le corpus de référence.

Le Tableau 5 montre les scores de nouveauté obtenus pour nos données. Les topics avec un score positif sont les topics T1,

T1	T2	T3	T4	T5	T6
108.81	1.56	-0.96	19.14	-0.61	-0.88

TABLE 5: Scores de nouveauté pour les topics de notre corpus.

T2 et T4 pour lesquels nous avons constaté (manuellement) qu'ils correspondent au nouveau sens *geste* de « quenelle ».

5. Explication des calculs : Le corpus *REF* contient 164 paragraphes dans le sens *culinaire* auxquels nous pouvons ajouter 31 paragraphes dans le sens *culinaire* issus du corpus *NOUV*, suite à notre vérification manuelle. Le corpus *NOUV* contient 342 paragraphes dont 38 n'ont pas le sens *geste* : 31 véhiculent le sens *culinaire* et 7 ont un thème indéterminé.

5 Conclusion

Nous avons présenté une application d'une méthode à base de Topic Models pour l'identification du nouveau sens du signifiant « quenelle ». Les diverses mesures proposées par Lau *et al.* (2012, 2014) constituent des critères pertinents dans notre cas. Cette méthodologie reste à valider et à reconduire sur d'autres changements de sens connus comme « caviar » ou « souris ». L'un des points problématiques de la méthode réside dans l'alignement des topics avec les définitions trouvées dans les dictionnaires : les répertoires de sens et la granularité sont fortement dépendants du dictionnaire utilisé. Il semblerait donc plus réaliste de détecter automatiquement les évolutions thématiques et de laisser aux lexicographes le soin de les décrire sur la base des indices automatiques.

Remerciements

Ces travaux ont été financés par l'Université de Strasbourg dans le cadre de l'Initiative d'Excellence (IdEx) 2012 (projet Logoscope).

Références

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**.

BOUSSIDAN A. & PLOUX S. (2011). Using topic salience and connotational drifts to detect candidates to semantic change. In *Proceedings of IWCS 2011*.

CABRÉ T. & NAZAR R. (2011). Towards a New Approach to the Study of Neology. In *Neology and Specialised Translation 4th Joint Seminar Organised by the CVC and Termisti*.

JI H., PLOUX S. & WEHRLI E. (2003). Lexical Knowledge Representation with Contexonyms. In *Proceedings of the 9th Machine Translation*, p. 194–201.

LAU J. H., COOK P., MCCARTHY D., GELLA S. & BALDWIN T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of ACL 2014*, Baltimore, USA.

LAU J. H., COOK P., MCCARTHY D., NEWMAN D. & BALDWIN T. (2012). Word sense induction for novel sense detection. In *Proceedings of EACL 2012*.

REUTENAUER C. (2012). *Vers un traitement automatique de la néosémie : approche textuelle et statistique*. PhD thesis, Université de Lorraine.

REUTENAUER C., JACQUEY E., LECOLLE M. & VALETTE M. (2010). Sémème au macroscope : genèse et variation sémiques d'une unité lexicale. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, Sapienza University of Rome : Sergio Bolasco, Isabella Chiari, Luca Giuliano.

ROMARY L., SALMON-ALT S. & FRANCOPOULO G. (2004). Standards going concrete : from LMF to Morphalou. In M. ZOCK, Ed., *COLING 2004 Enhancing and using electronic dictionaries*, Geneva, Switzerland.

STEYVERS M. & GRIFFITHS T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, **427**(7).

TEH Y. W., JORDAN M. I., BEAL M. J. & BLEI D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**(476).