Évaluation de la contextualisation de tweets

Patrice Bellot (1), Véronique Moriceau (2), Josiane Mothe (3), Eric SanJuan (4), Xavier Tannier (2)

(1) LSIS - Aix-Marseille Université (France)
(2) LIMSI-CNRS, University Paris-Sud (France)
(3) IRIT, UMR 5505, CNRS, Université de Toulouse, Institut Universitaire de Formation des Maîtres Midi-Pyrénées (France)
(4) LIA, Université d'Avignon et des Pays de Vaucluse (France)

patrice.bellot@univ-amu.fr {moriceau,xavier.tannier}@limsi.fr josiane.mothe@irit.fr eric.sanjuan@univ-avignon.fr

RÉSUMÉ. *Cet article s'intéresse à l'évaluation de la contextualisation de tweets. La contextualisation est définie comme un résumé permettant de remettre en contexte un texte qui, de par sa taille, ne contient pas l'ensemble des éléments qui permettent à un lecteur de comprendre tout ou partie de son contenu. Nous définissons un cadre d'évaluation pour la contextualisation de tweets généralisable à d'autres textes courts. Nous proposons une collection de référence ainsi que des mesures d'évaluation adhoc. Ce cadre d'évaluation a été expérimenté avec succès dans la contexte de la campagne INEX Tweet Contextualization. Au regard des résultats obtenus lors de cette campagne, nous discutons ici les mesures utilisées en lien avec les autres mesures de la littérature.*

ABSTRACT. *This paper deals with tweet contextualization evaluation. Text contextualization is defined as providing the reader with a summary allowing a reader to understand a short text that, because of its size is not self-contained. A general evaluation framework for tweet contextualization or other type of short texts is defined. We propose a collection benchmark as well as the appropriate evaluation measures. This framework has been experimented in INEX Tweet Contextualisation track. Based on the track results, we discuss these measures with regards to other measures from the litterature.*

MOTS-CLÉS : *Contextualisation, Évaluation, Résumé automatique, Informativité, Lisibilité*

KEYWORDS: *Contextualization, Evaluation, Automatic summarization, Informativeness, Readability*

1. Introduction

Nous définissons la contextualisation de textes courts comme la génération d'un résumé permettant de remettre en contexte un texte qui, de par sa taille, ne contient pas l'ensemble des éléments permettant à un lecteur de comprendre tout ou partie de son contenu. Les textes courts peuvent être de plusieurs natures : des requêtes soumises à un moteur de recherche, des SMS, des tweets, etc. Notre intérêt s'est plus particulièrement porté sur les tweets, à la fois pour leur popularité et parce que leur taille de 140 caractères maximum correspond bien à une situation dans laquelle la contextualisation peut être utile. Les informations échangées à travers des tweets concernent ainsi bien souvent des individus ou des évènements comme des élections, des conférences, des manifestations ou des tempêtes...

L'objectif de la contextualisation de textes courts est donc de fournir à son lecteur des informations qui rendent le message compréhensible. Si l'on considère par exemple le cas de la contextualisation de tweets, le tweet *"Bobby Brown – Fighting #WhitneyHouston's Family to See Bobbi Kristina"* pourrait être contextualisé par un résumé tel que celui de la Figure 1. Fournir un tel résumé implique de s'appuyer sur des ressources disponibles pour le constituer. Comme dans le cas général de la génération automatique de résumés, deux approches sont possibles : construire les résumés à partir de passages (par exemple des phrases) issus des textes des ressources originales (Goldstein *et al.,* 1999) ou paraphraser ces textes ressources (Genest *et al.,* 2010). La contextualisation de tweets peut ainsi être rapprochée de la génération de résumé à partir de documents retrouvés en réponse à une requête, tâche largement étudiée dans la littérature, en particulier dans les programmes TAC qui seront évoqués plus loin. Selon nous, elle s'en distingue par plusieurs aspects : les tweets sont limités en taille mais sont généralement plus qu'une succession de mots clés, ils s'insèrent dans un ou plusieurs flux qui ne se lisent pas toujours linéairement, emploient des *htags* faisant référence à des thèmes plus ou moins normalisés et ne peuvent souvent être compris qu'en faisant appel à l'historique et à des connaissances non explicites (le contexte). Nous souhaitons contribuer au développement des approches qui permettent d'éclairer le lecteur sur le contexte d'émission d'un tweet.

Dans cet article, notre contribution concerne plus spécifiquement la définition d'un cadre d'évaluation de la contextualisation de tweets. Nous proposons une collection de référence ainsi que des mesures d'évaluation *adhoc*. Nous discutons ces mesures eu égard à d'autres mesures de la littérature et des résultats de la campagne INEX Tweet Contextualisation qui a implémenté ce cadre d'évaluation.

Dans la section 2, nous présentons les travaux reliés. La contextualisation de texte est une nouvelle tâche de RI ; cependant, sur certains aspects, elle est proche des tâches de création de résumés automatiques. Nous nous focalisons donc sur les mesures d'évaluation du domaine. La section 3 présente les mesures que nous proposons pour évaluer le contenu et la lisibilité des résumés de contextualisation. La section 4 discute les mesures proposées en lien avec les autres mesures de la littérature et présente les résultats de la campagne INEX.

Whitney Elizabeth Houston (August 9, 1963 February 11, 2012) was an American recording artist, actress, producer, and model. Houston was one of the world's best-selling music artists, having sold over 170 million albums, singles and videos worldwide. Robert Barisford "Bobby" Brown (born February 5, 1969) is an American R&B singer-songwriter, occasional rapper, and dancer. After a three-year courtship, the two were married on July 18, 1992. On March 4, 1993, Houston gave birth to their daughter Bobbi Kristina Houston Brown, her only child, and his fourth. With the missed performances and weight loss, rumors about Houston using drugs with her husband circulated. Following fourteen years of marriage, Brown and Houston filed for legal separation in September 2006. Their divorce was finalized on April 24, 2007, with Houston receiving custody of their then-14-year-old daughter. On February 11, 2012, Houston was found unresponsive in suite 434 at the Beverly Hilton Hotel, submerged in the bathtub.

Figure 1. *Exemple d'un résumé contextualisant le tweet* "Bobby Brown – Fighting #WhitneyHouston's Family to See Bobbi Kristina". *Les phrases utilisées proviennent de Wikipedia.*

2. Évaluation de résumés

L'évaluation d'un résumé ne peut être entièrement automatisée car elle comprend deux dimensions : le contenu informatif du résumé et sa lisibilité. Ces deux dimensions ne sont pas indépendantes, il semble difficile d'évaluer le contenu informatif d'un texte totalement illisible et, de même, on ne s'intéresse pas à la lisibilité de manière abstraite, mais dans le contexte restreint d'un corpus de documents à résumer. Il est apparu possible d'évaluer automatiquement le contenu informatif d'un résumé en le comparant soit avec un ou plusieurs résumés de référence, soit directement avec les textes d'origine. Dans le cas d'utilisation de résumés de référence, ils sont alors généralement produits manuellement, mais lorsqu'ils sont construits, ils peuvent demeurer la référence, à la manière des jugements de pertinence produits dans les collections de référence en RI adhoc (les *qrels* de TREC par exemple). Ce type d'évaluation a été largement utilisé en particulier dans les campagnes d'évaluation DUC , puis TAC, organisées par le NIST.

L'évaluation des résultats produits peut se faire en estimant le pourcentage d'information des résumés de référence présent dans le résumé construit automatiquement, comme dans DUC 2003 avec l'outil SEE[1] ou selon des approches plus élaborées comme dans DUC à partir de 2004, sur la base des phrases les composant ou sur la base de n-grammes comme dans ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). L'évaluation pyramidale est une autre méthode d'évaluation (Nenkova *et al.,* 2004) où au lieu de comparer des distributions de n-grammes, on procède au préalable à l'identification de concepts clefs sur les résumés de référence

1. *http ://www.isi.edu/licensed-sw/SEE/*

et on évalue leur présence sous différentes formes dans les résumés produits automatiquement.

Ces modèles d'évaluation reposent cependant sur l'existence de résumés de référence par des humains. Le nombre de documents à résumer est donc limité et ne permet pas le passage à l'échelle.

Dans le cas d'un très grand nombre de documents à résumer, il semble naturel de trouver une mesure qui permette de comparer le contenu du résumé produit à celui de l'ensemble des documents dont doit être issu le résumé. Le plus intuitif est de comparer les distributions de mots ou de séquences de mots entre le résumé et les documents. Les mesures de divergence de Kullback-Leibler (KL) et de Jenssen-Shanon (JS) permettent de mesurer la similarité entre deux distributions de probabilité $P(i)$ et $Q(i)$.

3. Évaluation de la contextualisation de tweets

Dans cette section, nous définissons les mesures que nous proposons pour évaluer la contextualisation de tweets. Comme dans le cas des résumés automatiques, cette contextualisation doit être évaluée sur deux dimensions : le contenu et la lisibilité des résumés produits.

3.1. *Évaluation du contenu : la mesure LogSim*

Nous avons proposé un compromis entre ROUGE et KL. Il s'agit de la mesure LogSim qui comme ROUGE est orientée rappel sur la présence/absence de n-grammes possiblement à trous. Elle s'applique à comparer les distributions de fréquences entre un échantillon de passages de référence issus d'une très large collection de documents et les résumés produits par extraction de ces documents. Cette mesure introduite dans (SanJuan *et al.,* 2012) afin de mesurer la similarité de contenu d'un résumé avec un résumé de référence est robuste vis-à-vis de la variabilité des tailles autant des résumés que des références.

Soit T un ensemble de termes de référence et S l'ensemble de termes issus d'un résumé à évaluer. Nous considérons trois types de termes : les mots simples (unigrammes de ROUGE), les séquences de deux mots consécutifs adjacents (Bi-grammes de ROUGE) et les séquences de deux mots consécutifs éventuellement séparés par un ou deux mots (Bi-grammes à trous de ROUGE). On note par $P(t|X)$ la probabilité conditionnelle $\frac{f_X(t)}{f_X}$ avec X étant T ou S. $f_X(t)$ correspond à la fréquence du terme t dans l'ensemble X. La mesure LogSim que nous proposons est définie par :

$$LogSim(T,S) = \sum_{t \in T} P(t|T) \times \frac{\min(R(t,T), R(t,S))}{\max(R(t,T), R(t,S))} \quad [1]$$

$$R(t,X) = \log(1 + P(t|X) \times |T|) \quad [2]$$

LogSim a des propriétés identiques aux mesures de précision interpolée si la précision est définie comme le nombre de n-grammes dans le résumé de référence. Elle est définie sur des probabilités nulles et la fonction $log(1 + x)$ assure sa robustesse dans le cas des termes fréquents (mots outils, verbes communs) tout en ne pénalisant pas les termes de spécialité importants mais qu'il peut être difficile de capter lors de la construction d'un échantillon de passages pertinents. Par contre l'utilisation de cette même fonction rend LogSim incompatible avec un modèle LDA. LogSim est normalisée entre 0 et 1 du fait du facteur $P(t|T)$ mais elle n'est pas additive.

3.2. *Proposition pour l'évaluation de la lisibilité*

De notre point de vue, l'évaluation de la lisibilité et de la cohérence sémantique d'un texte nécessite une intervention humaine. Nous avons développé une interface Web à cet effet. Chaque résumé est constitué de passages et chaque passage est associé à quatre cases à cocher :

– syntaxe : indique que le passage contient un problème de syntaxe (mauvaise segmentation par exemple) ;

– anaphore : indique que le passage contient des références non résolues (p.e., impossibilité pour le lecteur de rattacher un pronom anaphorique à son antécédent) ;

– redondance : indique que le passage contient des informations déjà mentionnées dans les passages précédents ;

– à écarter : le passage est incohérent ou incompréhensible, même après avoir lu les passages précédents. Dans ce cas, les passages suivants sont évalués comme si ce passage n'avait pas été présent ;

– si le résumé est si mauvais que la lecture est arrêtée avant la fin, alors toutes les cases doivent être cochées à partir du dernier passage lu.

Le nombre des mots (jusqu'à 500) dans les passages valides ainsi que trois métriques sont ensuite utilisées pour l'évaluation de la lisibilité :

– métrique tolérante (fondée sur la pertinence) : un passage est considéré comme valide si la case "À écarter" n'a pas été cochée ;

– métrique intermédiaire : un passage est considéré comme valide si les cases "À écarter" et "Syntaxe" n'ont pas été cochées ;

– métrique stricte : un passage est considéré comme valide si aucune case n'a été cochée.

Pour chaque résumé, le texte des tweets (sans les *tags*) est affiché durant l'évaluation des passages rendant l'évaluation de la lisibilité elle-même contextuelle. De cette façon, des passages lisibles (sémantiquement interprétables en dehors de tout contexte) peuvent être écartés par les évaluateurs (car jugés non lisibles) à partir du moment où ils sont incompréhensibles étant donné le tweet d'origine.

4. Résultats et discussions

La collection utilisée pour l'évaluation dans campagnes INEX Tweet Contextualization 2011 et 2012 a été décrite dans (SanJuan *et al.,* 2012). Les documents sont issus d'une version récente et nettoyée de la Wikipedia en anglais. Les topics sont quant à eux des tweets sélectionnés par les évaluateurs, qui ont vérifié que la contextualisation était possible à partir de la Wikipedia, ainsi que d'environ 1000 tweets collectés automatiquement – ces derniers ne servant pas à l'évaluation finale. L'objectif était de s'assurer de la robustesse des systèmes, et d'éviter que les participants ne puissent réaliser des traitements manuels compte tenu de leur nombre.

4.1. *Comparaison de LogSim avec d'autres mesures*

Comme expliqué précédemment, dans le cadre d'une tâche de contextualisation d'un message court à partir d'un très grand nombre de documents, il serait beaucoup trop onéreux de constituer des résumés de référence établis par des humains. Sans ces résumés, nous ne pouvons pas calculer l'ordonnancement des résumés soumis avec ROUGE pour le comparer avec celui obtenu avec LogSim. Si à défaut des résumés pertinents nous choisissons de prendre l'échantillon de passages pertinents établis par les organisateurs, nous constatons une absence de corrélation statistique entre les deux ordonnancements. Nous constatons aussi une absence de corrélation avec les classements produits en utilisant KL et JS. Surtout, il se confirme que KL et JS sont très sensibles à la taille des résumés soumis et des échantillons produits de passages pertinents. Il est cependant possible de comparer LogSim aux mesures classiques de la RI appliquées aux entités. En effet, nous avons annoté dans le corpus tous les liens présents dans la Wikipedia. Ces liens correspondent à une annotation manuelle des termes devant faire référence à une autre entrée de l'encyclopédie collaborative, d'où la possibilité de les considérer comme une annotation manuelle des documents comparable à celle des SCUs utilisées dans la méthode pyramidale (section 2). Nous avons remplacé chaque passage de la référence par les entités présentes dans ce passage. Nous avons procédé de même pour toutes les soumissions des participants. Nous avons alors calculé précision, rappel et la F1-mesure (moyenne harmonique des deux) entre entités présentes dans chaque résumé produit et celles dans les références. Nous trouvons une forte corrélation statistique entre l'ordonnancement des systèmes participants produit avec LogSim sur les bigrammes (simples ou à trous) et celui induit par le F1-mesure calculée sur les entités (mesures de corrélation de Pearson et de Spearman toutes deux supérieures à 90 % avec une p-valeur très inférieure à 0,001).

4.2. *Discussion sur les résultats de la campagne*

Lors de la campagne 2011, 11 équipes avaient participé et soumis 23 runs. La campagne 2012 a quant à elle mobilisé 13 équipes de 10 pays différents et 33 runs ont été soumis. En 2011, 37 303 passages avaient été soumis alors qu'il y en a eu 671 191

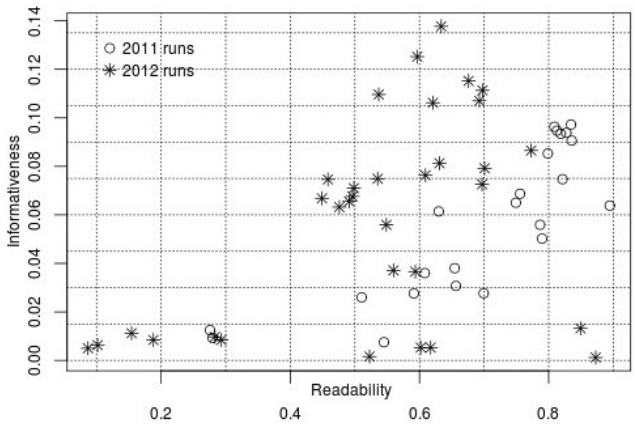

Figure 2. *Scores de lisibilité (Readability) et de contenu (Informativeness) pour les soumissions officielles de 2011 et 2012.*

en 2012. La figure 2 donne les résultats obtenus par les participants aux campagnes 2011 et 2012. Si la lisibilité n'a pas progressé, les scores pour le contenu ont nettement augmenté, malgré le fait qu'en 2011 les tweets du NYT étaient plus propres. Les différences de score de contenu supérieures à 1,5 % sont toutes significatives, tandis que pour la lisibilité, seules les différences de plus de 10 % sont significatives (t-test avec valeur $p < 0,05$).

Concernant les approches mises en œuvre, tous les participants (à l'exception de deux) ont utilisé des modèles de langue ; cependant, les systèmes qui se sont contentés d'un outil de recherche de passages ont un score inférieur à 5 %. Plusieurs participants ont reformulé les tweets à l'aide de la LDA (Latent Dirichlet Allocation) (Morchid *et al.,* 2012, Deveaud *et al.,* 2012) ou d'extraction terminologique (Vivaldi *et al.,* 2012), donnant de bons résultats (parmi les 10 meilleurs systèmes). Par ailleurs, tous les systèmes arrivés dans les cinq premiers ont utilisé un étiqueteur morpho-syntaxique (TreeTagger ou Stanford Core NLP). Vis à vis de la lisibilité, le meilleur système de 2011 (Ermakova *et al.,* 2012) a mis en œuvre de l'évaluation automatique de lisibilité et de la détection d'anaphore, ainsi que la prise en compte de la densité d'information des résumés. Pourtant ce système n'a pas été aussi efficace en 2012, probablement en raison des adaptations nécessaires à la variété des tweets. Les méthodes classiques de résumé automatique basées sur la phrase ont également montré de bons résultats, parmi les meilleurs scores (Crouch *et al.,* 2012, Deveaud *et al.,* 2012, Ganguly *et al.,* 2012). Le système s'étant classé deuxième en 2011 utilisait un algorithme de vote pour combiner plusieurs systèmes de résumé (Moreno *et al.,* 2012).

5. Conclusion

Le cadre d'évaluation que nous avons proposé et expérimenté dépasse le simple traitement des tweets. Compte tenu du nombre de systèmes participants indépendants supérieur à 30, du fait aussi que lors de l'édition 2012, les participants ont déclaré ne pas avoir utilisé le marquage des entités du corpus, nous avons montré ici que la mesure LogSim est corrélée à la présence/absence d'entités pertinentes dans les résultats. Cette propriété de LogSim permettrait d'étendre la tâche de contextualisation à d'autres ressources pour lesquelles on ne dispose pas d'annotation des entités. Ce cadre d'évaluation peut aussi s'étendre à un contexte multilingue à condition d'inclure suffisamment de participants susceptibles d'évaluer la réelle lisibilité des textes.

6. Bibliographie

, Forner, P., , Karlgren, J., , Womser-Hacker, C. (eds), *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, 2012.

Crouch C. J., Crouch D. B., Chittilla S., Nagalla S., Kulkarni S., Nawale S., « The 2012 INEX Snippet and Tweet Contextualization Tasks », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Deveaud R., Boudin F., « LIA/LINA at the INEX 2012 Tweet Contextualization track », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Ermakova L., Mothe J., « IRIT at INEX : Question Answering Task », *in* , S. Geva, , J. Kamps, , R. Schenkel (eds), *International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX)*, LNCS 7424, Springer, 2012.

Ganguly D., Leveling J., Jones G. J. F., « DCU@INEX-2012 : Exploring Sentence Retrieval for Tweet Contextualization », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Genest P.-E., Lapalme G., Yousfi-Monod M., « Jusqu'où peut-on aller avec des méthodes par extraction pour la rédaction de résumés ? », *Actes de TALN 2010*, 2010.

Goldstein J., Kantrowitz M., Mittal V., Carbonell J., « Summarizing text documents : sentence selection and evaluation metrics », *Proceedings of SIGIR'99*, 1999.

Lin C. Y., « ROUGE : A Package for Automatic Evaluation of Summaries », *Proceedings of the Workshop on Text Summarization Branches Out (WAS)*, July 25-26, 2004.

Morchid M., Linarès G., « INEX 2012 Benchmark a Semantic Space for Tweets Contextualization », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Moreno J. M. T., Velázquez-Morales P., « Two Statistical Summarizers at INEX 2012 Tweet Contextualization Track », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Nenkova A., Passonneau R., « Evaluating content selection in summarization : The pyramid method », *Proceedings of HLT-NAACL*, vol. 2004, 2004.

SanJuan E., Moriceau V., Tannier X., Bellot P., Mothe J., « Overview of the INEX 2012 Tweet Contextualization Track », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.

Vivaldi J., da Cunha I., « INEX Tweet Contextualization Track at CLEF 2012 : Query Reformulation using Terminological Patterns and Automatic Summarization », *in* Forner, Karlgren and Womser-Hacker (2012), 2012.