Mesures d'informativité et de lisibilité pour un cadre d'évaluation de la contextualisation de tweets

Patrice Bellot [1], **Véronique Moriceau** [2], **Josiane Mothe** [3], **Éric SanJuan** [4], **Xavier Tannier** [2]

1. *Aix-Marseille Université CNRS, LSIS UMR 7296, Marseille, France*
   patrice.bellot@univ-amu.fr
2. *LIMSI-CNRS, Université Paris-Sud, France*
   {veronique.moriceau,xavier.tannier}@limsi.fr
3. *IRIT, UMR 5505, CNRS, Université de Toulouse, ESPE Midi-Pyrénées, France*
   josiane.mothe@irit.fr
4. *LIA, Université d'Avignon et des Pays de Vaucluse, France*
   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 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 ad hoc. Ce cadre d'évaluation a été expérimenté avec succès dans le contexte de la campagne INEX Tweet Contextualization. Au regard des résultats obtenus lors de cette campagne, nous discutons ici les mesures proposées et les résultats obtenus par les participants.*

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. We discuss these measures and participants' results.*

MOTS-CLÉS : *contextualisation, évaluation, résumé automatique, informativité, lisibilité.*

KEYWORDS: *contextualization, evaluation, automatic summarization, informativeness, readability.*

DOI:10.3166/DN.15.1.55-73

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 pourraient 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. Les informations échangées à travers des tweets concernent bien souvent des individus ou des événements comme des élections, des conférences, des manifestations ou des catastrophes naturelles... Nous nous avons choisi d'étudier le cas des 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.

Ainsi, l'objectif de la contextualisation de textes courts est de fournir à son lecteur des informations qui rendent le message plus explicite. Par exemple, le tweet suivant :

Bobby Brown – Fighting #WhitneyHouston's Family to See Bobbi Kristina

pourrait être contextualisé par un texte synthétique (une sorte de résumé contextuel) tel que celui présenté dans la figure 1.

Un tel résumé contextuel est supposé fournir des informations complémentaires relatives au contenu du tweet, qui pourrait nécessiter une explication. Fournir un tel résumé de façon automatique 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 : par extraction où l'on construit les résumés en extrayant les passages les plus représentatifs (par exemple des phrases) des textes complets (Goldstein *et al.*, 1999) ou par synthèse où l'on génère des paraphrases de ces derniers (Genest *et al.*, 2010).

La contextualisation de tweets peut ainsi être rapprochée de la génération de résumés orientés par une requête où les documents à résumer sont trouvés par un moteur de recherche. Cette tâche est largement étudiée dans la littérature, en particulier dans les programmes TAC qui seront évoqués plus loin. Selon nous, elle se distingue de la contextualisation de tweets 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 *hashtags* 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). Dans cet article, nous souhaitons contribuer au développement des approches qui permettent d'éclairer le lecteur sur le contexte d'émission d'un tweet.

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. Nous discutons ces mesures au regard d'autres

mesures de la littérature et des résultats de la campagne INEX Tweet Contextualisation dans laquelle ce cadre d'évaluation est expérimenté depuis plusieurs années.

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 Wikipédia*

La suite de l'article est organisée comme suit. Dans la section 2, nous présentons les travaux reliés. Si la contextualisation de texte est une nouvelle tâche de Recherche d'Information (RI), elle est également proche des tâches de création de résumés automatiques. Nous nous focalisons ainsi plus spécifiquement sur les mesures d'évaluation du contenu informatif des résumés. Dans la section 3, nous présentons les données utilisées dans le cadre de l'évaluation que nous proposons en section 4. Nous proposons également dans la section 4 des mesures pour évaluer le contenu et la lisibilité des résumés. En particulier, la mesure d'informativité que nous proposons a l'avantage de ne pas être sensible à la taille du résumé produit. Enfin, la section 5 discute les mesures proposées en lien avec les autres mesures de la littérature et présente les résultats de l'évaluation INEX.

2. Évaluation de résumés : les mesures basées sur le contenu informatif

Les premiers travaux dans le domaine de la création automatique de résumés remontent aux années 1950. (Luhn, 1958) proposait alors d'utiliser la fréquence des mots et leur distribution pour estimer l'importance des mots puis des phrases afin de sélectionner celles qui apparaissaient dans le résumé produit automatiquement. Le même principe est utilisé dans la méthode MMR (*Maximal Marginal Relevance*) (Carbonell, Goldstein, 1998) qui sélectionne les phrases les plus pertinentes tout en minimisant la redondance. Il est également la base des travaux menés dans LexRank (Erkan, Radev, 2004) et TextRank (Mihalcea, Tarau, 2004) qui extraient les phrases considérées comme centrales ou dans (Ermakova, Mothe, 2012a) qui pondère différemment les éléments constitutifs des phrases (entités nommées, prépositions, ...) afin de choisir celles-ci.

Dans cette section, nous présentons les mesures couramment employées pour l'évaluation du contenu des résumés. L'évaluation d'un résumé comprend au moins deux dimensions fondamentales : le contenu informatif du résumé et sa qualité linguistique que nous rapprochons de sa lisibilité. Ces deux dimensions ne sont pas indépendantes : aussi informatif soit-il, au sens de la *quantité d'information*, le contenu informatif d'un texte totalement illisible (incompréhensible par un humain) ne devrait pas être pris en compte et, inversement, la lisibilité ne doit être considérée de manière isolée, mais dans le contexte restreint d'un corpus de documents à résumer. L'évaluation de ces deux facettes de la qualité d'un résumé ne semble pas pouvoir être entièrement automatisée. Nous nous intéressons ici à l'évaluation du contenu informatif ; l'évaluation de la lisibilité sera abordée en section 4.2.

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 et constituent une référence à la manière des jugements de pertinence produits dans les collections de référence en RI *ad hoc* (les *qrels* de TREC par exemple). Ce type d'évaluation a été largement utilisé en particulier dans les campagnes d'évaluation DUC (*Document Understanding Conferences* duc.nist.gov/) (Dang, 2005), puis TAC (*Text Analysis Conference*, www.nist.gov/tac/) (Dang, 2008), 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 avec ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004) également utilisée dans TAC. L'évaluation pyramidale est une autre méthode d'évaluation (Nenkova, Passonneau, 2004) dans laquelle, au lieu de comparer des distributions de n-grammes, on procède au préalable à l'identification manuelle de concepts clefs sur les résumés de référence ; leur présence est ensuite évaluée sous différentes formes dans les résumés produits automatiquement.

Ces modèles d'évaluation reposent cependant sur l'existence de résumés de référence produits 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. Des mesures utilisées à TAC consistent à comparer les distributions de mots ou de séquences de mots entre le résumé et les documents avec par exemple les mesures de divergence de Kullback-Leibler (KL) et de Jenssen-Shanon (JS).

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

Un point intéressant relatif aux mesures ROUGE, KL et JS est que l'on a montré leur corrélation avec une évaluation totalement humaine sous la condition de les appliquer à des résumés jugés lisibles. Sans cette condition, il est possible d'améliorer artificiellement les résultats obtenus avec ces mesures alors que le résumé produit n'aura pas de sens pour un humain (suite de termes par exemple).

Dans les sections suivantes, nous présentons les mesures Pyramide, ROUGE, KL et JS. Nous présenterons dans la section 4 la mesure plus adaptée que nous avons proposée pour évaluer la contextualisation de tweets dans le cas de résumés de taille variable. Des détails sur d'autres mesures d'évaluation de la RI se trouvent dans ce numéro spécial ainsi que dans (Pinel-Sauvagnat, Mothe, 2013).

2.1. L'évaluation par identification de pépites informationnelles

Dans le contexte de tâches de questions-réponses complexes, l'évaluation pyramidale a été introduite par (Nenkova, Passonneau, 2004) et étendue à l'évaluation de l'informativité de résumés. Cette méthode est basée sur la notion de SCU (*Summary Content Units*) ou pépites (Nuggets) définies à l'origine manuellement par des annotateurs. Il s'agit d'unités informationnelles regroupant différentes expressions qui correspondent au même contenu et qui sont pondérées selon le nombre d'annotateurs les ayant identifiées. Une expression, lorsqu'elle est sélectionnée pour faire partie d'un SCU y « contribue ». A chaque SCU est associé de façon manuelle un libellé qui exprime son contenu. Par ailleurs, à chaque SCU est associé un poids qui dépend du nombre de résumés de référence qui ont contribué au SCU. Ainsi, chaque unité est pondérée en fonction de sa popularité chez les annotateurs. Un résumé automatique sera d'autant bien noté qu'il contient des unités de fort poids. Cette méthode est par exemple utilisée pour évaluer la tâche de *Résumé d'opinions* introduite en 2008 dans la campagne TAC (Dang, 2008). Initialement, la construction des SCUs était donc manuelle et dépendante de l'accord inter-annotateur (Lin, Zhang, 2007). Les travaux récents en RI ont montré qu'elle pouvait être généralisée à la création automatique de références (Pavlu *et al*., 2012) sur la base d'extracteurs automatiques d'unités informationnelles plus générales (pépites ou nuggets) (Hennig *et al*., 2010 ; Ekstrand-Abueg *et al*., 2013). Par exemple, dans le cas de Wikipédia, on peut considérer comme pépites naturelles les liens internes qui renvoient à d'autres pages et qui contiennent souvent des reformulations de l'entité cible. Cependant, adopter a priori une telle approche comme mesure d'évaluation de l'informativité tend à réduire la tâche de création d'un résumé à celle d'une simple recherche et extraction d'entités. Ceci explique notre intérêt pour les mesures de comparaison de distributions de n-grammes exposées ci-après et indépendantes de la caractérisation de ces pépites informationnelles.

2.2. Les mesures ROUGE

ROUGE, pour *Recall-Oriented Understudy for Gisting Evaluation*, correspond à un ensemble de variantes de mesures qui comparent un résumé produit de façon automatique avec un (ou des) résumé(s) de référence généralement produit manuelle-

ment (Lin, 2004). Le principe général utilisé dans ROUGE est de compter le nombre d'items, par exemple des n-grammes, communs entre le résumé automatique et les résumés de référence. Une première série de variantes de ROUGE est la famille *ROUGE-n* qui compare les n-grammes des résumés comme suit (Lin, 2004) :

$$ROUGE\text{-}n = \frac{\sum_{S \in \{\text{résumés de référence}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{résumés de référence}\}} \sum_{gram_n \in S} Count(gram_n)}$$

avec :

– n la longueur des n-grammes considérés (ROUGE-1 et ROUGE-2 sont les mesures les plus communes),

– $Count_{match}(gram_n)$ le nombre maximum de n-grammes qui apparaissent à la fois dans le résumé à évaluer et dans les résumés de référence (cardinal de l'intersection la plus grande entre le résumé à évaluer et chacun des résumés de référence),

– $Count(gram_n)$ le nombre de n-grammes qui apparaissent dans les résumés de référence.

ROUGE-n est une mesure orientée rappel. ROUGE-1 est semblable à la mesure cosinus, mais pour $n > 1$, l'ordre des mots étant pris en compte, ROUGE est plus stricte que celle-ci (Alguliev *et al.*, 2011). ROUGE favorisera un résumé à évaluer qui contient des n-grammes qui apparaissent dans plusieurs références. Il faut noter que l'ajout de références différentes favorise des résumés variés puisque la valeur du dénominateur augmente avec le nombre de résumés de référence. Les variantes de ROUGE-n permettent d'affiner la capacité de la mesure à considérer les similarités des résumés par rapport à leurs n-grammes communs.

La mesure ROUGE-S (*Skip-Bigram Co-Occurrence Statistics*) permet de considérer des sauts entre les mots pour construire les n-grammes à comparer. Par exemple la chaine « a b c d » génère les bi-grammes suivant « a b » « a c » « a d » « b c » « b d » « c d ». Ces méthodes de construction des bi-grammes pénalisent des chaînes qui contiennent des mots communs mais dans des ordres différents. Le détail des variantes ROUGE est décrit dans (Lin, 2004). La mesure proposée en section 4 reprend ces notions de Skip-Bigram qui inclut les uni-grammes. La corrélation entre les évaluations manuelles et automatiques avec ROUGE sur des résumés varient de $0,80$ à $0,94$ (Louis, Nenkova, 2009) d'où leur adoption lors des campagnes de résumés automatiques.

Cependant ROUGE n'est appropriée que si l'on cherche à comparer des résumés humains de très bonne qualité (les références) à des résumés produits automatiquement (Saggion *et al.*, 2010). Elle est donc adaptée à l'évaluation d'une tâche de traitement automatique des langues à échelle humaine. Les campagnes DUC puis TAC utilisant ROUGE ne portaient au maximum que sur une cinquantaine de documents. Les évaluateurs devaient lire l'ensemble de ces documents et produire un résumé aussi exhaustif que possible d'une taille prédéfinie au mot près. Pour que l'évaluation soit stable avec ROUGE, il est nécessaire de produire au moins cinq résumés de référence

établis par différents experts (Louis, Nenkova, 2009). Il n'est donc pas possible d'utiliser ROUGE pour évaluer le contenu informatif d'un document pour des résumés de différente taille, ou lorsque la production des résumés de référence est trop coûteuse comme dans le cas de documents très longs ou d'un très grand nombre de documents à résumer, ce qui est le paradigme de la RI.

2.3. Divergence de Kullback-Leibler et de Jenssen-Shanon

Dans le cas d'un très grand nombre de documents à résumer, il semblait 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 (Moriceau *et al.*, 2009). 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 $Q(i)$. La mesure de KL est définie par :

$$D_{KL}(P||Q) = \sum_i \ln \frac{P(i)}{Q(i)} P(i)$$

La mesure de JS est une variation symétrique et lissée de KL et est définie par :

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M)$$

avec

$$M = \frac{1}{2}(P + Q)$$

JS est la mesure qui a été montrée comme ayant la plus forte corrélation avec ROUGE (Louis, Nenkova, 2009) mais cette mesure est sensible à la taille des résumés. Elle est donc difficile à appliquer dans le cadre d'une tâche de recherche d'information ciblée où la taille des résumés produits peut varier en fonction de la quantité d'information pertinente trouvée dans la collection de documents. KL n'est pas défini pour des probabilités nulles, il est alors nécessaire d'introduire des techniques de lissage propre aux modèles de langage.

Si les mesures ROUGE sont peut-être trop sensibles à la présence/absence de séquences de mots communs entre résumé et référence, KL a l'inconvénient opposé de ne pas permettre de distinguer entre une très faible fréquence et une absence totale. Par ailleurs, KL et JS ne permettent que d'évaluer des résumés génériques et sont difficiles à adapter au cas de résumés guidés par une requête ou un sujet, comme cela est le cas dans le cadre de la contextualisation de textes courts. Pour toutefois pouvoir l'appliquer, il serait nécessaire de constituer une référence de passages pertinents et d'appliquer ces mesures entre la référence ainsi produite et les résumés proposés. Cette référence serait constituée d'extraits de documents et ne serait pas un résumé comme dans le cas de ROUGE. La difficulté d'adapter KL et JS à ce cas de figure

est qu'il y a une grande variabilité dans la taille des références elles-mêmes. Lorsqu'il y a peu de passages pertinents dans les documents, l'estimation des probabilités par lissage devient très délicate (SanJuan *et al.*, 2010).

Dans les sections suivantes, nous présentons la collection que nous avons définie dans le cadre de l'évaluation de la contextualisation de tweets. Nous présenterons également une nouvelle mesure qui lève les limites que nous venons de détailler.

3. Une collection pour évaluer la contextualisation de tweets

Nous présentons ici les données utilisables pour évaluer la contextualisation de tweets telles que définies dans les campagnes d'évaluation INEX Tweet Contextualization 2011, 2012 et 2013. Elles sont disponibles pour tous les participants à INEX [2] et sont décrites avec précision [3] (voir aussi (SanJuan *et al.*, 2012)).

3.1. Contenu de la collection de test

La méthodologie d'une campagne d'évaluation se doit de permettre également l'évaluation de systèmes *a posteriori*. Durant une campagne, les participants produisent des réponses aux tâches définies et ces réponses sont évaluées. Après la campagne, de nouvelles productions de résultats doivent pouvoir être proposées et évaluées. Pour cela, la collection de référence doit se baser sur des éléments de référence et des mesures calculables automatiquement. Ainsi, une collection de référence pour évaluer une tâche de contextualisation de tweets doit contenir :

– un ensemble de documents de référence : ce sont ces documents qui sont utilisés comme ressource pour extraire des éléments constitutifs d'un résumé correspondant à la contextualisation d'un tweet ;

– un ensemble de tweets à contextualiser : ils correspondent aux besoins (ou *topics*) usuels des campagnes d'évaluation ;

– des résumés, ou des passages de référence qui permettront de comparer leurs contenus aux résumés produits ;

– des mesures d'évaluation, dont au moins une partie doit permettre une évaluation *a posteriori*.

3.2. Documents

La collection de documents servant de source à la contextualisation de tweets a été construite sur la base d'une version (*dump*) de Wikipédia en anglais de novembre 2011 puis en novembre 2012. Pour être réaliste et permettre de répliquer les résultats, cette collection a été figée de sorte que les documents soient antérieurs aux tweets.

2. https://inex.mmci.uni-saarland.de/people/register.jsp
3. https://inex.mmci.uni-saarland.de/tracks/qa/

Nous avons « nettoyé » les documents afin d'obtenir des documents XML, de manière à faciliter les traitements des participants. Nous avons ainsi supprimé les notes et les références bibliographiques, plus difficiles à exploiter. Par ailleurs, nous n'avons gardé que les pages non vides, c'est-à-dire qui contiennent au moins une section. Les documents résultants sont composés d'un titre (`title`), d'un résumé (`a`) et de sections (`s`). Chaque section possède un sous-titre (`h`). Le résumé et les sections sont composés de paragraphes (`p`) et chaque paragraphe peut contenir des entités (`t`) qui font référence à d'autres pages Wikipédia. La figure 3 présente un exemple d'un document. La DTD des documents de la collection est la suivante :

```
<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT a (p+)>
<!ELEMENT s (h, p+)>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA | t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>
```

Figure 2. DTD des documents de la collection

3.3. Les tweets "topics"

En 2011, nous avions proposé des tweets à partir des titres d'articles du New York Times (NYT). En 2012 et 2013, afin d'être plus réaliste, nous avons préféré construire une collection de tweets réels. Ainsi, nous avons collecté manuellement des tweets et choisi 63 tweets en 2012 et 70 tweets en 2013, de telle façon que :

– les tweets contiennent un contenu « informatif » ; nous n'avons sélectionné que des tweets issus de comptes non personnels (par exemple @CNN, @TennisTweets, @PeopleMag, @science...) ;

– les documents de la collection de référence contiennent des contenus en lien avec le tweet, de sorte que la contextualisation soit possible.

Par exemple, le tweet suivant est considéré comme pertinent :

```
Very cool! An interactive animation of van Gogh's
"The Starry Night." http://t.co/ErJCPObh (thanks
@juliaxgulia)
```

```
<?xml version="1.0" encoding="utf-8"?>
<page>
<ID>2001246</ID>
<title>Alvin Langdon Coburn</title>
<s o="1">
 <h>Childhood (1882-1899)</h>
  <p o="1">Coburn was born on June 11, 1882, at 134 East
    Springfield Street in <t>Boston, Massachusetts</t>,
    to a middle-class family. His father, who had
    established the successful firm of Coburn &amp;
    Whitman Shirts, died when he was seven.
    [...] </p>
  <p o="2">In 1890 the family visited his maternal
    uncles in Los Angeles, and they gave him a 4 x 5
    Kodak camera. He immediately fell in love with
    the camera, and within a few years he had developed
    a remarkable talent for both visual composition
    and technical proficiency in the <t>darkroom</t>.
    [...]</p>
    [...]
</page>
```

Figure 3. Extrait d'une page Wikipédia

alors que le tweet suivant n'est pas retenu car non informatif et non contextualisable *via* Wikipédia :

```
Mom: "All you do is sit on that computer all day!" Me:
"Lies! I sit on the chair."
```

Au-delà des tweets collectés et contrôlés manuellement, nous avons constitué une collection de tweets plus importante basée sur une collecte automatique. Cet ensemble de tweets est composé à la fois des tweets contrôlés et de tweets « tout venant », en provenance toutefois des mêmes comptes Twitter prédéfinis. 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 d'un trop faible nombre de tweets à traiter. Par ailleurs, le fait d'avoir plusieurs types de tweets peut également être valorisé dans des expérimentations variées. La collection est donc prévue pour n'être évaluée que sur les tweets contrôlés, mais lors de leur participation initiale, les participants ignoraient quels étaient ces tweets. Les tweets sont fournis à la fois au format texte, sans métadonnées, et au format JSON avec les métadonnées associées. La figure 4 donne un exemple de tweet au format JSON.

```
"created_at":"Wed, 15 Feb 2012 23:32:22 +0000",
"from_user":"FOXBroadcasting",
"from_user_id":16537989,
"from_user_name":"FOX Broadcasting",
"id":169927058904985600,
"text":"Tensions are at an all-time high as the
@AmericanIdol
Hollywood Round continues, Tonight at 8/7c. #Idol",
"to_user":null,
(...)
```


Figure 4. Extrait d'un tweet de la collection

3.4. Échantillon de passages de référence

En 2012, nous avons constitué des échantillons de passages de référence en évaluant les résumés proposés par les participants pour chacun des 63 tweets collectés manuellement. Pour chaque tweet, nous avons conservé les 60 meilleurs passages de chaque participant en nous appuyant sur leur score RSV (*Relevance Score Value*) que les systèmes devaient fournir et qui permet de classer les passages par ordre de pertinence par rapport à un tweet. Les doublons ont ensuite été supprimés puis les passages ont été présentés aux évaluateurs par ordre alphabétique. Ainsi, chaque passage a été évalué indépendamment des autres. La structure et la lisibilité des résumés n'ont pas été évaluées à ce stade. Les évaluateurs n'avaient qu'à donner un jugement binaire sur la pertinence du passage vis-à-vis du tweet. Au final, nous avons évalué 16 754 passages issus de 33 soumissions valides réalisées par 13 équipes distinctes. Dans cet échantillon 2 801 passages ont été jugés pertinents. Le nombre moyen de passages par tweet est de 265,9 (55,1 pour les passages jugés pertinents). La longueur moyenne d'un passage est de 30,03 tokens.

La même démarche a été appliquée pour les tweets collectés en 2013. Nous avons évalué 21 046 passages issus de 24 soumissions valides réalisées par 11 équipes distinctes. Dans cet échantillon, 2 254 passages ont été jugés pertinents. Le nombre moyen de passages par tweet est de 350,77 (40,98 pour les passages jugés pertinents). La longueur moyenne d'un passage est de 36,53 tokens.

En 2011, la même méthode avait été appliquée mais nous disposions aussi des articles du NYT correspondant aux tweets. Ces articles avaient été utilisés comme résumés de référence pour une seconde évaluation automatique.

4. 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.

4.1. 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. La quantité $f_X(t)$ 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. LogSim est normalisée entre 0 et 1 du fait du facteur $P(t|T)$ mais elle n'est pas additive.

4.2. Proposition pour l'évaluation de la lisibilité

L'évaluation des résumés produits automatiquement s'appuie sur la notion de cohésion (continuité et progression sémantique et référentielle), de cohérence locale (au niveau des phrases) ou globale (au niveau du résumé complet) et de lisibilité générale. L'évaluation d'un résumé peut se baser sur le fait qu'il corresponde ou non à un texte cohérent et qu'il contienne les concepts clefs (Jones, 2007). L'identification des concepts clefs peut s'appuyer sur l'utilisation d'un ensemble de questions qui doivent trouver leur réponse dans le ou les textes originaux. Les questions doivent également

trouver leurs réponses dans les résumés automatiques produits pour que celui-ci soit considéré comme pertinent. Ce peut être des questions de compréhension de texte (Morris *et al*., 1992) ou des questions relatives au contenu du texte initial (Mani *et al*., 2002).

De notre point de vue, l'évaluation de la lisibilité et de la cohérence sémantique d'un texte nécessite une intervention humaine. Ainsi, selon notre approche, chaque résumé est constitué de passages et chaque passage est évalué selon quatre critères à satisfaire inspirés de (Pitler, Nenkova, 2008), (Feng *et al*., 2010) et des critères de cohérence et cohésion définis par (Halliday, Hasan, 1976) :

– **lisibilité syntaxique** : ce critère est satisfait si le passage ne contient pas de problème de syntaxe (contre exemple : mauvaise segmentation) ;

– **lisibilité anaphorique** : ce critère est satisfait si le passage ne contient pas de références non résolues (contre exemple : impossibilité pour le lecteur de rattacher un pronom anaphorique à son antécédent) ;

– **absence de redondance** : ce critère est satisfait si le passage ne contient pas d'informations déjà mentionnées dans les passages précédents ;

– **cohérence** : ce critère est satisfait si le passage est cohérent et compréhensible après avoir lu les passages précédents. Dans le cas contraire, les passages suivants sont évalués comme si ce passage n'avait pas été présent.

Si un résumé est si mauvais que la lecture est arrêtée avant la fin, alors aucun des critères n'est satisfait à partir du dernier passage lu.

Pour l'évaluation de la lisibilité, trois métriques sont utilisées ainsi que le nombre des mots (jusqu'à 500) dans les passages considérés comme valides selon ces métriques :

– métrique tolérante (fondée sur la pertinence) : un passage est considéré comme valide si le critère de « cohérence » est satisfait ;

– métrique intermédiaire : un passage est considéré comme valide si les critères de « cohérence » et « lisibilité syntaxique » sont satisfaits ;

– métrique stricte : un passage est considéré comme valide si tous les critères sont satisfaits.

Durant l'évaluation des passages, chaque résumé est présenté avec le texte du tweet (sans les marques de *hashtags* mais en laissant son nom) 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.

Afin de faciliter l'évaluation de la lisibilité, nous avons mis en place une interface web qui permet aux évaluateurs de cocher des cases lorsque les critères sont satisfaits.

5. Résultats et discussions

5.1. 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, la mesure ROUGE ne peut pas être appliquée pour calculer l'ordonnancement des résumés soumis et comparer cet ordonnancement 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 fondées sur l'extraction de pépites informationnelles telles que Pyramide en considérant les liens présents dans la Wikipédia. 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 2.1). Nous avons alors calculé les scores de précision, rappel et F1-mesure (moyenne harmonique des deux mesures précédentes) 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 bi-grammes (simples ou à trous) et celui induit par la 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). Ainsi, si l'on dispose d'une référence de passages pertinents, les deux mesures sont similaires. En effet, les entités de la Wikipédia semblent suffisantes pour discriminer entre passages pertinents et non pertinents tandis que la mesure LogSim reste sensible à la présence de pépites informationnelles implicites.

5.2. Discussion sur les résultats de la campagne

Lors de la campagne 2011, 11 équipes avaient participé et soumis 23 runs. La campagne 2012 a mobilisé 13 équipes de 10 pays différents et 33 runs ont été soumis. En 2013, 11 équipes de 9 pays ont soumis 24 runs.

En 2011, 37 303 passages avaient été soumis alors qu'il y en a eu 671 191 en 2012 et 210 460 en 2013. La figure 5 donne les résultats obtenus par les participants aux trois campagnes de 2011 à 2013. Si la lisibilité n'a pas progressé en 2012, 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 $p < 0{,}05$). En 2013, les scores pour le contenu et

la lisibilité étaient très fortement corrélés (Kendall test: $\tau > 90\,\%$, $p < 0{,}001$), ce qui montre que les systèmes ont intégré ces deux dimensions.

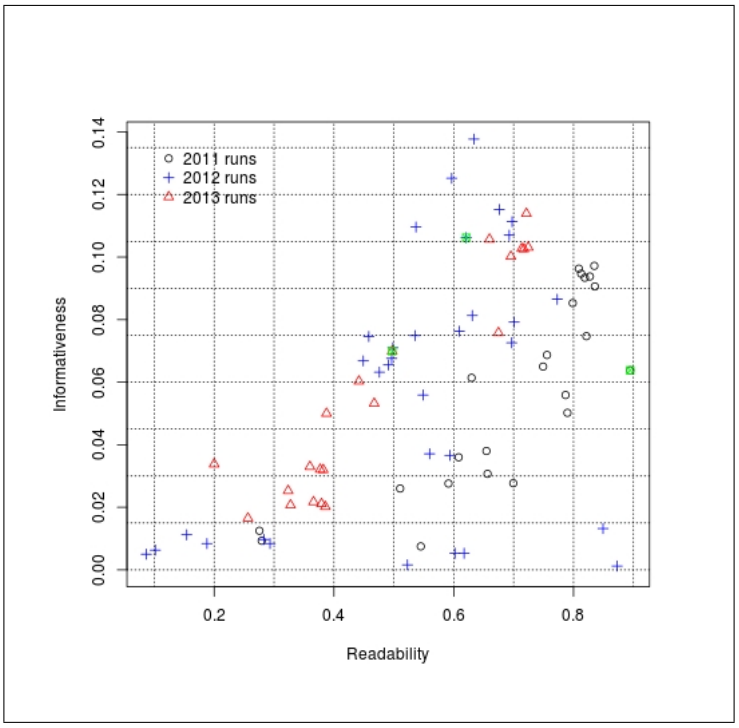

Figure 5. Scores de lisibilité (Readability) et de contenu (Informativeness) pour les soumissions officielles de 2011 à 2013

Concernant les approches mises en œuvre, la plupart des participants 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 du modèle LDA (Morchid, Linarès, 2012) ou d'extraction terminologique (Vivaldi, Cunha, 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). En 2013, le système ayant obtenu le meilleur score pour le contenu a mis en œuvre un traitement spécifique des *hashtags* (Deveaud, Boudin, 2013). Pour la campagne 2014, nous avons proposé un ensemble de tweets à contextualiser où chaque tweet est associé à une entité de Wikipédia afin de donner aux participants un point d'entrée dans la Wikipédia et de ne pas trop pénaliser les systèmes qui ne traitent pas les *hashtags*.

Vis-à-vis de la lisibilité, le meilleur système sur les trois années (Ermakova, Mothe, 2012b) 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 sélection de phrases ont également montré de bons résultats, parmi les meilleurs scores (Crouch *et al*., 2012 ; Deveaud, Boudin, 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, Velázquez-Morales, 2012).

6. Conclusion

Le cadre d'évaluation que nous avons proposé et expérimenté dépasse le simple traitement des tweets. Il s'agit à notre connaissance d'une des premières tâches qui requiert la combinaison effective d'approches de RI et de traitement automatique de la langue naturelle écrite (TAL). Les mesures originales que nous proposons ont pour le contenu, une interprétation en termes de précision et de rappel, et pour la lisibilité intègrent une évaluation qualitative avec un questionnaire qui incite à l'emploi de techniques de TAL avancées.

L'évaluation de l'informativité au sens de la présence d'unités informationnelles univoques est en théorie orthogonale à l'évaluation de la lisibilité hors contexte, c'est-à-dire en faisant abstraction de la requête. Cependant, au cours de cette campagne nous avons été amenés à évaluer l'informativité contenue dans des passages alors que l'appréciation de cette informativité requiert que le passage soit lisible. De même, l'évaluation de la lisibilité hors contexte ne fait pas sens. En effet, tout extrait de la Wikipédia est lisible mais il faut que cet extrait réponde à la requête et dans ce cas, qu'il soit un minimum informatif. Il en résulte que nos axes d'évaluation ne sont pas indépendants et la question de leur unification se pose.

Dans les travaux futurs, nous souhaitons étudier la corrélation entre la présence ou l'absence d'entités pertinentes dans les résumés soumis et la mesure LogSim que nous avons proposée.

Le cadre d'évaluation que nous avons proposé peut aussi s'étendre à un contexte multilingue à condition d'inclure suffisamment de participants susceptibles d'évaluer la réelle lisibilité des textes.

Remerciements

Nous souhaitons remercier l'Agence Nationale de la Recherche qui a participé au financement des recherches présentées ici, au travers du projet CAAS-Contextual Analysis and Adaptive Search, ANR-10-CORD-001-01.

Bibliographie

Alguliev R. M., Aliguliyev R. M., Hajirahimova M. S., Mehdiyev C. A. (2011). MCMR: Maximum coverage and minimum redundant text summarization model. *Expert Syst. Appl.*, vol. 38, n° 12, p. 14514-14522.

Carbonell J., Goldstein J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM*

SIGIR conference on Research and development in information retrieval, p. 335-336. New York, NY, USA, ACM. http://doi.acm.org/10.1145/290941.291025

Crouch C. J., Crouch D. B., Chittilla S., Nagalla S., Kulkarni S., Nawale S. (2012). The 2012 INEX Snippet and Tweet Contextualization Tasks. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop)*.

Dang H. T. (2005). Overview of duc 2005. In *Proceedings of the document understanding conference*.

Dang H. T. (2008). Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Proceedings of the First Text Analysis Conference*.

Deveaud R., Boudin F. (2012). LIA/LINA at the INEX 2012 Tweet Contextualization track. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop)*.

Deveaud R., Boudin F. (2013). Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization. In *CLEF (Online Working Notes/Labs/Workshop)*.

Ekstrand-Abueg M., Pavlu V., Aslam J. A. (2013). Live nuggets extractor: a semi-automated system for text extraction and test collection creation. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, T. Sakai (Eds.), *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, p. 1087-1088. ACM.

Erkan G., Radev D. R. (2004, décembre). LexRank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, vol. 22, n° 1, p. 457–479. http://dl.acm.org/citation.cfm?id=1622487.1622501

Ermakova L., Mothe J. (2012a). IRIT at INEX 2012: Tweet Contextualization. In *CLEF Online Working Notes/Labs/Workshop*.

Ermakova L., Mothe J. (2012b). 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)*. Springer.

Feng L., Jansche M., Huenerfauth M., Elhadad N. (2010). A Comparison of Features for Automatic Readability Assessment. In *Proceedings of COLING 2010, Poster Volume*.

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

Ganguly D., Leveling J., Jones G. J. F. (2012). DCU@INEX-2012: Exploring Sentence Retrieval for Tweet Contextualization. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop)*.

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

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

Halliday M., Hasan R. (1976). *Cohesion in English*. London: Longman.

Hennig L., D. L. E. W., S. A. (2010). Learning Summary Content Units with Topic Modeling. In *Proceeding of COLING 2010, Poster Volume*, p. 391-399.

Jones K. S. (2007). Automatic summarising: The state of the art. *Information Processing Management*, vol. 43, n° 6, p. 1449 - 1481. http://www.sciencedirect.com/science/article/pii/S0306457307000878 (<ce:title>Text Summarization</ce:title>)

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

Lin J., Zhang P. (2007). Deconstructing Nuggets: the Stability and Reliability of Complex Question Answering Evaluation. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), *Proceedings of the 30th International ACM SIGIR conference on research and development in Information Retrieval*, p. 327-334. ACM.

Louis A., Nenkova A. (2009). Performance Confidence Estimation for Automatic Summarization. In *Proceedings of EACL*, p. 541-548. The Association for Computer Linguistics.

Luhn H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal*, vol. 2, p. 159-165.

Mani I., Klein G., House D., Hirschman L., Firmin T., Sundheim B. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering*, vol. 8, n° 1.

Mihalcea R., Tarau P. (2004, July). TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Morchid M., Linarès G. (2012). INEX 2012 Benchmark a Semantic Space for Tweets Contextualization. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop)*.

Moreno J. M. T., Velázquez-Morales P. (2012). Two Statistical Summarizers at INEX 2012 Tweet Contextualization Track. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop)*.

Moriceau V., SanJuan E., Tannier X., Bellot P. (2009). Overview of the 2009 QA Track: Towards a Common Task for QA, Focused IR and Automatic Summarization Systems. In S. Geva, J. Kamps, A. Trotman (Eds.), *Proceedings of INEX*, vol. 6203. Springer.

Morris A. H., Kasper G. M., Adams D. A. (1992). The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research*, vol. 3, n° 1, p. 17-35.

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

Pavlu V., Rajput S., Golbus P. B., Aslam J. A. (2012). IR system evaluation using nugget-based test collections. In E. Adar, J. Teevan, E. Agichtein, Y. Maarek (Eds.), *Proceedings of the fifth ACM international conference on Web Search and Data Msagining (WSDM)*, p. 393-402. ACM.

Pinel-Sauvagnat K., Mothe J. (2013). Mesures de la qualité des systèmes de recherche d'information. *Ingénierie des Systèmes d'Information*, vol. 18, n° 3, p. 11-38.

Pitler E., Nenkova A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Saggion H., Torres-Moreno J.-M., Cunha I. da, SanJuan E., Velázquez-Morales P. (2010). Multilingual Summarization Evaluation without Human Models. In C.-R. Huang, D. Jurafsky (Eds.), *Proceedings of COLING (Posters)*, p. 1059-1067. Chinese Information Processing Society of China.

SanJuan E., Bellot P., Moriceau V., Tannier X. (2010). Overview of the INEX 2010 Question Answering Track (QA@INEX). In S. Geva, J. Kamps, R. Schenkel, A. Trotman (Eds.), *Proceedings of INEX*, vol. 6932, p. 269-281. Springer.

SanJuan E., Moriceau V., Tannier X., Bellot P., Mothe J. (2012). Overview of the INEX 2012 Tweet Contextualization Track. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop).*

Vivaldi J., Cunha I. da. (2012). INEX Tweet Contextualization Track at CLEF 2012: Query Reformulation using Terminological Patterns and Automatic Summarization. In P. Forner, J. Karlgren, C. Womser-Hacker (Eds.), *CLEF (Online Working Notes/Labs/Workshop).*