Quality of life and breast cancer: medical information retrieval in internet health forums

Thomas Opitz

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)
Institut de Mathématiques et de Modélisation de Montpellier (I3M)
Université Montpellier 2

Workshop ≪ Evaluation et Analyse de la Qualité de Vie en Oncologie ≫

ICM, Montpellier
03/04/2014

The project

i3m Institut de Mathématiques et de Modélisation de Montpellier

Christian Lavergne
Cyrille Joutard

Thomas Opitz

LIRMM Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

Sandra Bringay
Jérôme Azé

cnrs

um2 UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES

ICM Institut régional du Cancer Montpellier | Val d'Aurelle

Caroline Mollevi

Quality of life (QoL) as a clinical concept

- **multidimensional**

  physical – psychological – social

- definition and **quantification** of QoL-dimensions is **difficult**

- **auto-questionnaires** : EORTC–QLQ–C30 [1]
  breast cancer specific module QLQ–BR23

1. [Aaronson et al., 1993]

Quality of life in social media

- **social media** $\rightsquigarrow$ user empowerment, knowledge democratization
- **large textual databases**

  `Cancerdusein.org` forum : 675 members, 17,000 posts.

- **anonymous communication**
  $\rightsquigarrow$ easier expression of affective states (emotions, opinions, doubts, ...) [2]

Our project's objectives

extract **clinically relevant information**

- items of the QLQ–BR23 questionnaire
- online health forums (`Cancerdusein.org`)
- compare to traditional clinical studies

2. [Hancock et al., 2007]

Main challenges

- texts are **heterogeneous and noisy**

  $\rightsquigarrow$ standard preprocessing methods

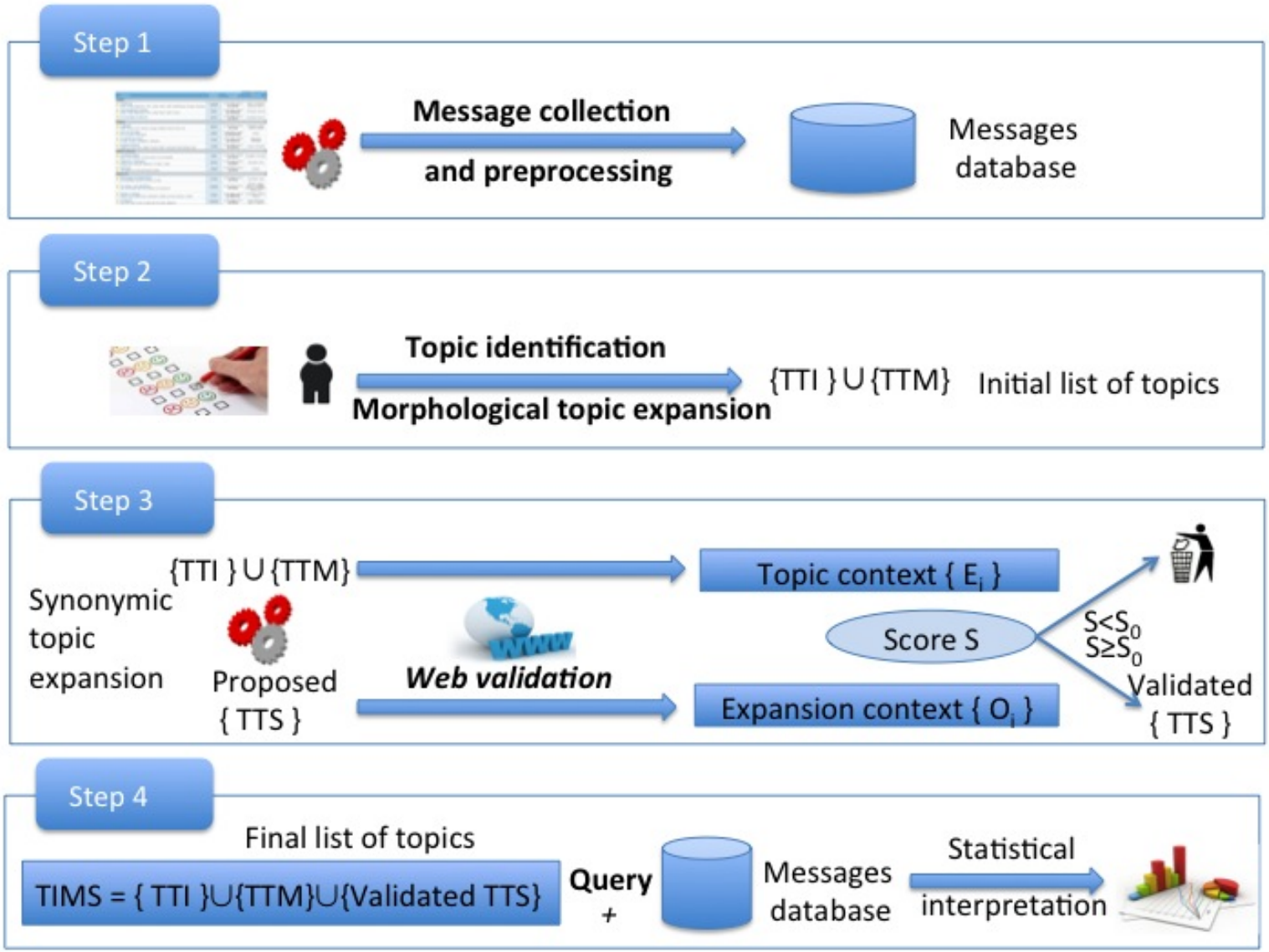
- many lexical variations of a given topic keyword

  $\rightsquigarrow$ **query expansion** techniques to **enrich topic descriptions**

- need **user metadata** $\rightarrow$ socio-economic data, clinical history, ...

  $\rightsquigarrow$ retrieve information from data

Overview of our approach



Step 1 : Data preprocessing

1. **spell correction**
2. **dimension reduction**
    1. removal of stopwords like *et*, *de*, *ne*, ...
    2. lemmatization [3] : *traite*, *traitant*, *traité*, *traitées*, ... $\rightsquigarrow$ *traiter*

Example :

moi j ai faita un reconstruction avec expendeur et je suis satisfaite du resultat
1) *moi* **j'***ai fait_ un reconstruction avec exp***a***ndeur et je suis satisfaite du r***é***sultat*
2.1) *j'ai fait reconstruction expandeur satisfaite résultat*
2.2) *je faire reconstruction expandeur satisfaire résultat*

3. `treetagger`

Step 2 : Manual definition of topics

QLQ-BR23

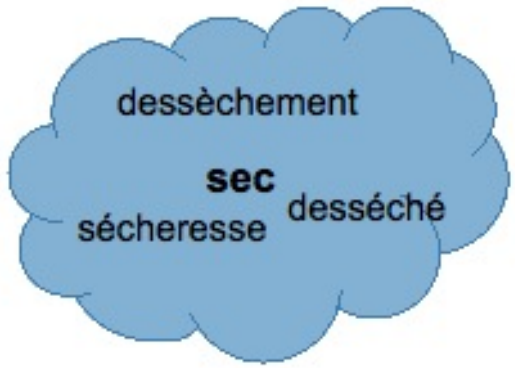
I1 : Avez-vous eu la bouche sèche?
I2 : La nourriture ou la boisson … ?
...

secondary effects of the treatment (e.g. hair loss)
symptoms in the breast region
body image
sexuality
...

identify topics → keyword representation

Topic 1 = {bouche sec}
Topic 2 = {boisson goût, nourriture goût}
...

morphological expansion



Topic 1 = {bouche sec, **buccal sec, bouche dessèchement, bouche desséché ...**}
Topic 2 = {boisson goût, nourriture goût, **aliment goût …**}
...

Step 3 : Automatic synonymic topic expansion

lexical variations beyond morphology $\rightsquigarrow$ many false negatives

> *"Mauvais goût dans la bouche et* **langue pâteuse** *ah ma bouche réel point faible ! ! !"*

> *"[...] car ma* **langue** *faisait* **bizarre** *et avait un bouton [...]"*

use **synonym expressions** for topic keywords

synonym tools : `www.synonymo.fr` and `www.cnrtl.fr/synonymie`

⚠ polysemy

- synonymy is context-sensitive
- synonym terms too numerous for manual selection

$\rightsquigarrow$ define a **score** for **automatic ranking** of synonym expansion
$\rightsquigarrow$ keep only useful expansions

Defining contexts

Context = frequent co-occurrences

- **synonym context** must be **close to the original context**
- context defined from a large text corpus (`yahoo.fr`)
  - unigrams/bigrams from 40 search engine snippets for each keyword
  - feature weight = frequency of co-occurrence

Bouche sèche - Xérostomie - Doctissimo
www.doctissimo.fr/html/sante/encyclopedie/bouche-seche.htm En cache
Avoir une **bouche sèche**, cela peut arriver à n'importe qui ponctuellement. Mais quand cela persiste, il ne faut pas se contenter de boire pour hydrater sa **bouche**.

Bouche sèche - Traitement et symptômes de la sécheresse ...
www.santepratique.fr/bouche-seche.php En cache
La **bouche sèche** pourrait être prévenue en partie par une hygiène rigoureuse mais non agressive de la cavité buccale et une hydratation suffisante.

...

topic : {*bouche*|*buccal*|... *sec*|*dessécher*|...}

topic context : *bouche sec (0.68), buccal (0.58), sécheresse (0.45), sec (0.42), bouche (0.25), mauvais haleine (0.2), xérostomie (0.17), symptôme (0.13), salive (0.13), ...*

---

langue seche et tres blanche - Troubles ORL - FORUM Santé
forum.doctissimo.fr/sante/troubles-orl/langue-seche... En cache
bonjour voila depuis plus de six mois j'ai ma **langue** de plus en plus blanche, tres blanche et **seche** , j'ai consulté mon generaliste il m'a donne des anti bio, et c ...

Bouche **sèche** la nuit / le matin? - Yahoo Questions/Réponses
fr.answers.yahoo.com/question/index?qid=20090714173539... En cache
"Bouche **sèche** la nuit / le matin?" ... ça fait plusieurs années que j'ai la bouche **sèche** (parfois même la **langue** collé au palais) ...

...

expansion : *langue sec*

expansion context : *langue (0.95), langue sec (0.8), sec (0.72), bouche sec (0.4), mauvais haleine (0.2), bouche (0.19), sensation (0.18), ...*

we define the **chi-square score**

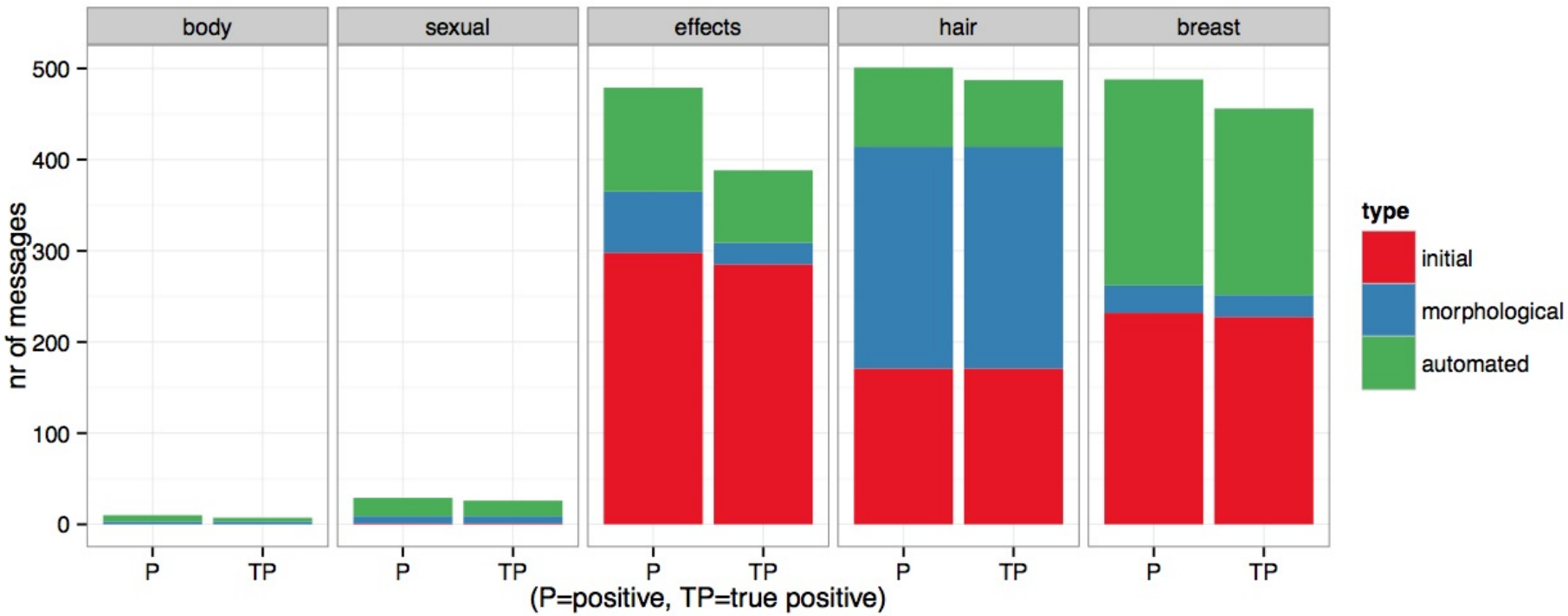
$$S = 1 - \left(\sum_{i=1}^{k} E_i \right)^{-1} \times \sum_{i=1}^{k} \max(E_i - O_i, 0)^2 / E_i \quad \in [0, 1].$$

- E_i – **weights** of the most important **topic** co-occurrences
- O_i – corresponding **weights** for the **expansion**

$\rightsquigarrow$ disjoint contexts for $S = 0$ and identical contexts for $S = 1$

Results

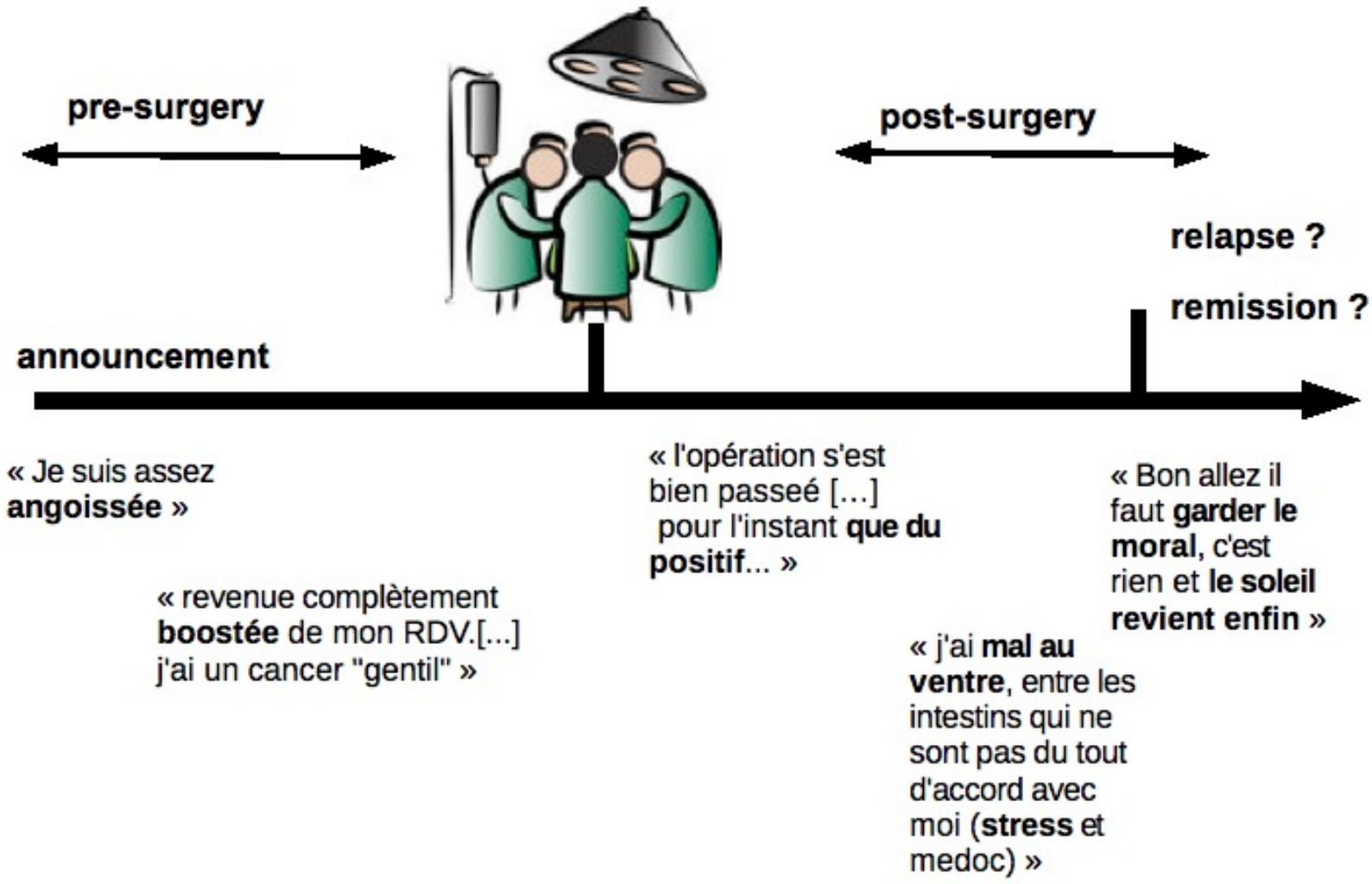
- keep automatic synonymic expansions if $S \geq 0.2$, manual validation for $0.1 \leq S \leq 0.2$
- 5 topic categories



Clinical evaluation of results

further information is necessary

- **clinical history of forum users** $\rightsquigarrow$ **supervised classification methods**
- information about quality of life related to occurrences
  $\rightsquigarrow$ **semi-automatic sentiment analysis** in free texts



Conclusions

- patients discuss **clinically relevant issues in health forums**
- **semi-automatic approach** works
- close **cooperation** between **data experts** and **clinical experts** :
  - **information** in social media is **abundant**
  - which information is **useful ?**
  - how can we **extract** it ?

Further reading

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleishman, S. B., de Haes, J. C., et al. (1993).
QLQ-C30 : a quality-of-life instrument for use in international clinical trials in oncology.
Journal of the National Cancer Institute, 85(5) :365–376.

Akay, A., Dragomir, A., and Erlandsson, B.-E. (2013).
A novel data-mining platform leveraging social media to monitor outcomes of Januvia.
In *35th Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, pages 7484–7487. IEEE.

Balahur, A. (2013).
Sentiment analysis in social media texts.
In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128.

Hancock, J. T., Toma, C., and Ellison, N. (2007).
The truth about lying in online dating profiles.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 449–452. ACM.

Subirats, L., Ceccaroni, L., Lopez-Blazquez, R., Miralles, F., García-Rudolph, A., and Tormos, J. M. (2013).
Circles of health : Towards an advanced social network about disabilities of neurological origin.
Journal of Biomedical Informatics, 46(6) :1006–1029.