Automatic acquisition of synonyms from French UMLS for enhanced search of EHRs

Natalia Grabar\textsuperscript{1}, Paul-Christophe Varoutas\textsuperscript{2}, Philippe Rizand\textsuperscript{2}, Alain Livartowski\textsuperscript{2}, Thierry Hamon\textsuperscript{3}

\textsuperscript{1} Université Paris Descartes, Faculté de Médecine; Inserm, U872; SPIM, 75006 Paris; HEGP, Paris
\textsuperscript{2} Institut Curie, Département d’Information Médicale, 25 rue d’Ulm, 75005 Paris, France
\textsuperscript{3} LIPN, UMR 7030, CNRS, Université Paris 13, Villetaneuse

Outline

- Introduction
- Method
- Material
- Experiment and Evaluation
- Conclusions and Perspectives
Prométhée (1/2)

Prométhée, a framework at the Institut Curie, designed for integration:

- Electronic health records (EHR)
- Heterogeneous biomedical data

Prométhée enables several levels of integration:

- Aggregation
- Cross-interrogation
- Visualization
- Statistical analysis
Prométhée (2/2)

- Deployed in the hospital since 2002
- Daily used by healthcare professionals
- Full-text search engine:
  - based on key-words
- Challenge for users:
  - guess the *right* key-words in order to access the right data
- Challenge for the NLP people (*i.e.*, this work):
  \[\rightarrow\] *provide linguistic resources for indexing and query expansion*
  - adapted to the clinical context
Variety of lexical resources

- **Morpho-syntactic resources: most complete**
  - General language
    - tools for POS-tagging and lemmatisation (Brill, Treetagger, Flemm)
  - specific databases:
    - Celex (Burnage, 1990), MorTal (Hathout et al, 2001)
  - Biomedical domain: UMLS Specialized Lexicon, similar resources for other languages (UMLF)

- **Synonymous relations: few available**
  - General language: WordNet (Fellbaum, 1998)
  - Biomedical domain: ongoing? (Smith et al, 2004)

- *Providing resources = acquisition + enrichment of resources*
Objectives

Propose a method for acquisition of synonym relations
Exploit semantic relations between words based on relations
between terms

- term compositionality
- syntactic invariants of terms

\{\textit{grippe aviaire}, \textit{influenza aviaire}\} \rightarrow \{\textit{grippe}, \textit{influenza}\}
\{\textit{gastric contents}, \textit{stomach contents}\} \rightarrow \{\textit{gastric}, \textit{stomach}\}
Observations on compositionality of biomedical terms

- Within the biomedical area, terms are sometimes coined on the same scheme
  - C0016627 *original* synonyms:
    - *grippe aviaire*
    - *influenza aviaire*
    - *peste aviaire*

- It is possible to induce the set of *elementary* synonyms:
  - grippe, influenza, peste

⇒ Generalize this observation:
- exploiting the compositional structure of terms
- using existing structured terminologies
Acquisition of synonym relations (1/4)

Compositionality: the meaning of a complex expression is fully determined by its syntactic structure, the meaning of its parts and the composition function (Partee, 1984)

Hypothesis:

- Compositionality preserves synonymy within complex terms
- Meaning $M$ of two complex terms $A \ rel \ B$ and $A' \ rel \ B$:
  
  $$M(A \ rel \ B) = f(M(A), M(B), M(\text{rel}))$$
  
  $$M(A' \ rel \ B) = f(M(A'), M(B), M(\text{rel}))$$

- If terms $A \ rel \ B$ and $A' \ rel \ B$ are synonymous
  
  $\implies$ Synonymy between their components $A$ and $A'$ can be deduced

  $\implies$ Syntactic analysis of terms

N Grabar et al.
Acquisition of synonym relations (2/4)

Rule 1: expansion

If both terms are synonymous and their expansion components are identical

\{grippe aviaire, influenza aviaire\} (avian influenza)

\[ \Rightarrow \{grippe, influenza\} (influenza) \]
Acquisition of synonym relations (3/4)

Rule 2: head

If both terms are synonymous and their head components are identical
\{contenu gastrique, contenu stomacal\} \{gastric contents, stomach contents\}

\[\implies \{gastrique, stomacal\} \{gastric, stomach\}\]
Acquisition of synonym relations \((4/4)\)

Rule 3: head and/or expansion

If both terms are synonymous and either their head components or expansion components are synonymous

\(\{\text{angine pustuleuse, pharyngite vésiculeuse}\} \{\text{aphtous pharyngitis, vesicular pharyngitis}\}\)

\(\{\text{angine, pharyngite}\} \text{ are already known synonyms}\)

\[\{\text{pustuleux, vésiculeux}\} \{\text{aphtous, vesicular}\}\]
Preprocessing of terminology: NLP platform
Providing syntactic analysis of terms

Ogmios (Hamon et al, 2007):
- Recognition of named entities: TagEN (Berroyer, 2004)
- Segmentation into words and sentences
- POS-tagging, lemmatisation: Genia (Tsuruoka et al, 2005)
  \[\text{angines/NP} \Rightarrow \text{angine/NS}\]
  \[\text{pustuleuse/AF} \Rightarrow \text{pustuleux/AM}\]
- Syntactic parsing of terms: YATEA (Aubin et al, 2006)
  Noun phrases:
  Adj Noun, Noun Prep Noun, Adj Adj Noun, ...
  \[\text{angine/N pustuleuse/A}\]
  Building the syntactic trees of terms
Evaluation of results

- Manual validation of the inferred elementary relations
  - each inferred pair is examined
  - source series of synonyms are examined
  ➞ accuracy of the inferred pairs

- Comparison of the inferred resource with available resources synonyms
  - *le Petit Robert*
  - *Masson*
  - French UMLS
  ➞ overlap between these sets of synonyms
  ➞ coverage of the inferred synonyms
Material

- **French subset of the UMLS**
  - 71,883 concepts: 156,404 terms
  - C0016627: *grippe aviaire, influenza aviaire, peste aviaire*
  \[\rightarrow\] Original resource for inferring elementary synonym relations

- **Two more sets of synonyms:**
  - general language synonyms *le Petit Robert*
    - 140,141 pairs
  - medical language synonyms *Masson*
    - 831 pairs
  \[\rightarrow\] Comparison and evaluation purposes
Preprocessing of terminology

- 156,404 terms (71,883 concepts) are fully parsed through the NLP platform:
  - segmentation
  - POS-tagging
  - lemmatising
  - syntactic parsing

- 76,240 original synonym terms $\Rightarrow$ 54,058 synonym pairs used for inducing the elementary synonym relations
Acquisition of synonym relations

- 1,196 pairs of inferred elementary synonym relations
- Three inferred pairs have large support (number of original series of synonyms):
  - 14: \{affection, maladie\} \{affection, disease\}
  - 12: \{maladie, syndrome\} \{disease, syndrome\}
  - 10: \{cancer, tumeur maligne\} \{cancer, malignant tumor\}
- Remaining elementary synonyms: small support
  often single pairs of original synonyms
Linguistic typology of the induced synonym pairs

- Orthographic variants:
  \{acathisie, akarhisie\}, \{embolie, embole\} (*embolus*)
- Abbreviations:
  \{biop, biopsie\}, \{EEG, électro-encéphalogramme\}
- Ellipse:
  \{insuffisance artérielle, insuffisance\} \{artery insufficiency, insufficiency\}
- Named entities:
  \{Bartholin, Duverney\}, \{Saint Jean, rhumatismale\}
- Scientific vs. popular words:
  \{maladie, pathologie\} \{disease, pathology\},
  \{abcès, empyème\} \{abscess, empyema\}
- Morphologically related words:
  \{vermiculaire, vermiforme\} (*vermiform*)
- No common formal features (most of the induced synonyms):
  \{grave, sévère\} *severe*,
  \{cancer, tumeur maligne\} \{cancer, malignant tumor\}
Evaluation (1/2)

- Manual evaluation of the 1,196 generated pairs:
  - performed by a computational scientist
  - correct: 99.3% (n=1,188)
  - rejected: 0.08% (n=1)
  - undecided: 0.6% (n=7)

⇒ High precision of the proposed method
  - acquisition performed on controlled terminological data
  - inferring rules strongly exploit syntactic invariants of terms

- Compositionality principle:
  - exploitable for lexical resource acquisition
Evaluation (2/2)

- **Erroneous pair:**
  - C0038814: \{coup de soleil, sensibilité au soleil\} \{sunburn, solar sensitiveness\}
  - *coup de soleil*: idiomatic expression

- **Borderline pairs:**
  - pairs with named entities (26)
    - glandes de Bartholin
    - glandes de Duverney
    - glandes vulvovaginales
    - glandes vestibulaires majeures
  - pairs with ellipses (102)
    - \{acide chlorhydrique gastrique, acide gastrique\}
    - \{activation granulocyte neutrophile, activation neutrophile\}
Comparison with available resources

- **Directly available synonym pairs within UMLS (36)**
  - \{tumeur maligne, cancer\} \{malignant tumor, cancer\}
  - \{saignement, hemorragie\} \{bleeding, hemorrhage\}
  \[\Rightarrow\] Detection of "hidden" synonyms

- **Common pairs with Masson medical dictionary (2)**
  - \{neurinome, neurilemmome\} \{neuroma, neurilemmoma\}
  - \{rétrocontrôle, feedback\} \{feedback\}
  \[\Rightarrow\] Not satisfying coverage

- **Common pairs with le Petit Robert dictionary (105)**
  - \{cancérologie, oncologie\} \{cancerology, oncology\}
  - \{contour, forme\} \{contour, shape\}
  - \{faux, imaginaire\} \{false, imaginary\}
  - \{ponction, aspiration\} \{punction, suction\}
  \[\Rightarrow\] Common-language resource should be adapted
Conclusion

- **Method for inferring elementary synonymous relations:**
  - Exploiting compositionality principle
  - Applying three rules based on syntactic dependency analysis
  - Based on the use of structured terminologies
  - Language-independent

- **Application to the French UMLS:**
  - High-quality results: over 99% precision
  - Possible weak points:
    - named entities, ellipses, idiomatic expressions
  - Efficiency of the NLP tools

- **Comparison with the available resource of synonyms:**
  - Very partial overlap with existing synonyms
    - Detection of new synonyms
    - Enrichment of available synonyms
Perspectives

- Using corpora and the inferred synonyms for enriching and extending terminologies, lexica, ...
  - Institut Curie: using EHR content for extending the ICD-O

- Using these resources for document indexing and query expansion

- Testing their efficiency with other biomedical terminologies

- Applying the method to other languages as far as:
  1. the required linguistic processing can be realised
  2. synonym relations between complex terms are available