Using Lexical Knowledge Patterns for Terminology Work in English and French: Some Important Differences

Elizabeth Marshman
University of Ottawa / Observatoire de linguistique Sens-Texte
elizabeth.marshman@uottawa.ca

16th European Symposium on Language for Special Purposes
August 2007

Presentation Outline

- Introduction
- Motivation and goals
- Methodology
- Results
- Conclusions
  - Suggestions for future work

Introduction

- Conceptual relations: relations between concepts in a concept system
  - Help to describe, delimit and differentiate between concepts
    - GENERIC-SPECIFIC, PART-WHOLE, CAUSE-EFFECT…
    - ASSOCIATION, DISEASE-SYMPTOM…
- "While names of relations are few and little used, there are ways of *expressing* relations that are part of the average speaker's everyday vocabulary." (Chaffin and Hermann 1988)

Introduction (2)

- Knowledge patterns
  - “[L]anguage combinations that frequently identify a particular conceptual relation…. For example, patterns such as *X is a kind of Y, an X is a Y, As include Bs, Cs and Ds,* indicate generic-specific relations.” (Meyer 1994)

Introduction (3)

- Lexical knowledge patterns
  - Prototypical structure: X + marker + Y
    - X, Y are expressions of concepts (e.g., terms, usually N/NP)
    - marker is a lexical unit or series of lexical units indicating a relationship between them
  - Cause-effect: e.g., X leads to Y
    Arteriosclerosis leads to strokes.
  - Association: e.g., X is linked to Y
    High cholesterol is linked to heart disease.

Tools for Terminologists

- Knowledge patterns can indicate *knowledge-rich contexts* (KRCs) in text corpora (Meyer 2001)
  - Segments of texts that provide at least one piece of information that is useful for conceptual analysis
    - Generally a conceptual relation
- Computer tools programmed with knowledge patterns can analyze corpora semi-automatically
  - Locate candidate KRCs
  - Present them to a user (e.g., terminologist) for interpretation and use

Tools for Terminologists (2)

- Retrieved candidate KRCs may be processed to identify pertinent information, sort contexts
  - By relation/sub-relation present
  - By expressions of concepts linked by relation

- Pattern may be described more or less specifically
  - Character string representing marker < Regular expression specifying marker/marker form, (form of) expression of related concepts, structure(s) in which these appear

➷ Savings of time and effort for the user

Motivation for the research

- Approach used successfully in many languages
    - But generally only one at a time
- Terminology work is largely bi-/multilingual
    - Need for tools that can process corpora in different languages in parallel
- Question: Will the approach work the same way in different languages?

General goals

- To observe and describe occurrences of lexical knowledge patterns for CAUSE–EFFECT and ASSOCIATION relations in English and French medical texts
- To compare these observations in English and French to evaluate pertinent similarities and differences
- To evaluate the impact these may have on possibilities for developing semi-automatic pattern-based KRC extraction tools for terminology work

The relations

- **CAUSE-EFFECT**
  - Barrière (2002)
- CREATION
  - e.g., *X induces Y*
  - e.g., *X conduit à Y*
- DESTRUCTION
  - e.g., *X anti-Y* (EN, FR)
- MAINTENANCE
  - e.g., *X required for Y*
  - e.g., *X permet Y*
- PREVENTION
  - e.g., *X prevents Y*
  - e.g., *X suppresseur de Y*
- MODIFICATION
  - e.g., *effect of X on Y*
  - e.g., *X module Y*
- INCREASE
  - e.g., *X promotes Y*
  - e.g., *X favorise Y*
- DECREASE
  - e.g., *X reduces Y*
  - e.g., *X inhibe Y*
- PRESERVATION
  - e.g., *X sustains Y*
  - e.g., *X limite Y*

The relations

- ASSOCIATION: significant co-occurrence of factors
  - *X is associated with Y*
  - *X est lié à Y*
    - RISK
      - *X is a risk factor for Y*
      - *X est un facteur de risque de Y*
    - CORRELATION
      - *X correlates with Y*
      - *X est corrélé avec Y*

  - NB Association is often a precursor of observations of causal relations, but not sufficient evidence to prove them; the relations are different but closely linked

The corpora

- Domains: Breast cancer, Heart disease
  - Etiology, development, effects, diagnosis, treatment, prevention
- Texts: mostly specialized journal articles
  - Small proportion of popularized articles
- Corpus size:
  - English: ±575,000 tokens
  - French: ±700,000 tokens

The data

- 15 candidate terms in each language, representing four semantic classes
  - e.g., *chemotherapy, cell, c-reactive protein, pathogenesis, atherosclerosis*
  - e.g., *chimiothérapie, cellule, cholestérol, coagulation, athérosclérose*
- which were used to extract, with WordSmith Tools, ± 1,400 contexts in each language

The data

- which when manually analyzed produced 442 English and 349 French relation occurrences
  - ± 20-30% ASSOCIATION and 70-80% CAUSE-EFFECT
- which included 154 English and 167 French candidate relation markers
  - ± 20% ASSOCIATION and 80% CAUSE-EFFECT

The analysis

- Number of relation occurrences
- Markers
  - Number of distinct markers observed
  - Proportions of relation occurrences associated with each marker
- Marker and pattern forms
  - Variation in marker form
  - Variation in expressions of concepts linked by markers
- Quantitative and qualitative evaluations

Results of the research

- Similarities noted in many factors analyzed
  - Both quantitative and qualitative
  - Promising for future development of bilingual applications
- However, differences in certain factors suggest differences in performance
  - Suggests language-specific development, refinement of approaches advisable

Number of relation occurrences

- Success of approach relies on expression of relations by means of knowledge patterns
- In this sample from the two corpora, proportions of contexts containing knowledge patterns expressing relations higher in English
  - English= 3.19 contexts per occurrence
  - French= 3.99 contexts per occurrence
- Suggests density of occurrences would be interesting to evaluate in more targeted study
  - Related to terms, to corpus texts?
  - Potential difference in expression of relations using these types of patterns, indicating need for larger corpora or other strategies in French?

Markers: Representation

- Any lexical knowledge pattern-based application requires formal representation of markers
  - With more or less specificity
  - With more or less definition of their environment
- Candidate KRCs with inadequately represented markers would be missed
- Aspects:
  - Marker variety
  - Marker form

Marker variety

- Ratio of number of analyzed occurrences to number of markers identified consistently higher in English than in French data
  - Overall: **2.87** in English; **2.09** in French
  - Cause-effect: **2.62** in English; **2.04** in French
  - Association: **3.79** in English; **2.33** in French

  - i.e., wider variety of markers in smaller number of occurrences in French

Marker variety (2)

- Number of markers required to reach a given proportion of occurrences observed
  - Most frequently observed markers likely to be targeted as good candidates for use
  - Cause-effect
    - 50%: **17** most frequent in English, **30** most frequent in French
    - 75%: **46** most frequent in English, **73** most frequent in French
  - Association almost identical in two languages
    - 50%: **6** most frequent in English and French
    - 75%: **12** most frequent in English and **13** most frequent in French

Marker variety (3)

- More markers may be required in French to retrieve similar numbers of candidate KRCs
  - Though could be interesting to evaluate inter-relation difference
- Increase in numbers of markers required accompanied by increase in complexity of developing applications in French

Marker form

- Variation in form of marker components, their order (excluding purely inflectional variation)
  - e.g., *role of X in Y, X's role in Y, X (plays a) role in Y, role for X in Y…*

- Complicates representation of markers in pattern forms for use
- Can lead to exclusion of KRCs if markers inadequately represented

- Aspects:
  - Passive forms of verbal markers
  - Interruption of complex markers

Passive forms of verbal markers

- Change in marker form
  - Often accompanied by addition of marker elements (e.g., *by*, *par*) that change pattern structure
  - e.g., X causes Y; Y is caused by X
- Need for adjustment of pattern forms
  - including inversion of X and Y in structure if these are to be identified automatically
- Significantly more occurrences of verbal markers in passive forms in English data
  - English: **14%** (24/175)
  - French: **4%** (5/140)

Interruptions of complex markers

- Significantly more occurrences with interruption of markers by expression of a related concept in English
  - e.g., *link between X and Y, effect of X on Y*
  - English: **12%** (54/442)
  - French: **4.5%** (16/349)

- Markers in both corpora often interrupted by other elements external to pattern
  - However, proportionally higher in French

Marker form (2)

- Factors in combination strongly suggest increased complexity of representing marker forms in English
  - Accompanied by increased complexity of developing pattern-based applications
  - Potential loss of some candidate KRCs if pattern and marker forms do not account for variation
  - Difficulty of further processing contexts increased
- Some qualitative differences may increase variation in performance between languages

Form of expressions of concepts

- Representing expressions of related concepts essential for more specifically defined pattern forms, generally in automatic identification of expressions of related concepts
  - Prototypical structure = NP + marker + NP
  - Divergence from prototypical term forms in representation of patterns may lead to loss of candidate KRCs
- Aspects:
  - Anaphora
  - Non-nominal expressions of related concepts

Anaphora

- Affect both form and content of contexts
  - Design of pattern forms that excludes forms such as pronouns may lead to loss of candidate KRCs
  - Contexts retrieved may not be complete or interpretable out of larger context

Anaphora (2)

- Anaphora significantly more frequent in French data
  - e.g., *Celles-ci/elles/ces protéines provoquent Y*

- Contexts containing anaphoric expressions replacing expressions of concepts:
  - English: **6%** (27/442)
  - French: **11%** (37/349)

Non-nominal expressions of concepts

- Specific pattern forms that specify the form of expressions of related concepts likely to exclude candidate KRCs with different types of these expressions
- Information may require processing for further applications
  - e.g., construction of ontologies, linking term records

Non-nominal expressions (2)

- Significantly more frequent in French overall
  - **8%** (37) in English; **17%** (58) in French

- Higher proportions of occurrences of non-nominal elements of all types in French
  - pronouns, e.g., *Ceci cause Y*
  - adjectives, e.g., *X est un facteur de risque cardiovasculaire*
  - clauses, e.g., *X1 fait X2, contribuant à Y*

Non-nominal expressions (3)

POS (of head)	**English**	**Frenc h**
n.	92%	83%
pron.	2% (10)	5% (17)
adj.	3% (15)	6% (21)
v.	1% (4)	1% (5)
claus e	2% (8)	4% (15)

- Significantly more contexts containing pronouns, clauses/verbs
- Trend towards significance in higher proportion of adjectives

Form of expressions of concepts (2)

- Differences indicate increased challenges in creating French pattern forms
  - Likely to increase difficulty in creating pattern-based applications
  - More candidate KRCs may be excluded by variations in form if patterns do not account for them
- Potential for increased challenges in using information present in French

Conclusions

- Similarities in the prevalence and nature of several factors show promise for developing bilingual pattern-based computer tools for identifying KRCs.
- However, in other areas, tools likely to function differently in the two languages
  - Approaches and expectations require adaptations
  - Improving results likely to require targeting specific issues in each language, for each project

Conclusions (2)

- Results in marker variety indicate a need to develop patterns for more markers in French to retrieve contexts
- Results observed in both languages indicate challenges in developing specifically defined pattern forms
  - Challenges in English data linked to marker form
  - Challenges in French data more linked to expression of related concepts

Conclusions (3)

- Types of applications affected will vary
  - Marker variety: all
  - Marker form: especially patterns involving strictly defined marker forms
  - Expression of related elements:
    - Tools that use strictly defined pattern structures including expressions of related concepts
    - Tools that automate identification of these expressions and ultimately concepts

Future work

- Need to study:
  - Pertinent issues in-depth, in detail
    - In specifically designed study for analysis of a given phenomenon
  - Issues in specific applications
  - Combined effects of phenomena
- Sub-analyses
  - by text type, level of specialization
    - possibly grade of evidence
  - by sub-domain
  - by relation
- Further evaluation:
  - of more relations
  - in more languages

Thank you! Any questions?

Thanks are extended to Marie-Claude L'Homme and Sylvie Vandaele. The generous support of the Social Sciences and Humanities Research Council and the Université de Montréal is also gratefully acknowledged.