Bilingual Terminology Mining

GUIDE: PROF. PUSHPAK BHATTACHARYYA

BY: MUNISH MINIA (07D05016)
PRIYANK SHARMA (07D05017)
Content

- **Introduction**
- **Multilingual Terminology Mining Chain**
  - Term Extraction
  - Term Alignment
    - Direct Context-Vector Method
    - Translation of Lexical Units
- **Linguistic Resources**
  - Comparable Corpora
  - Bilingual Dictionary
- **Conclusion**
Introduction

- Text mining research generally adopts “big is beautiful” approach
  - Justified by the need of large amount of data in order to make use of statistic or stochastic methods\(^1\)
- Hypothesis: The quality rather than the quantity of the corpus matters more in terminology mining
- Web is used as a Comparable Corpus
- The comparability of the corpus should not only be based on the domain or the sub-domain, but also on the type of discourse

\(^1\): Manning and Schütze, 1999
Multi lingual Terminology Mining Chain

- **Architecture:**

  - **Source language documents**
  - **Target language documents**
  - **WEB**
  - **Terminology Extraction**
  - **Lexical alignment process**
  - **Bilingual Dictionary**
  - **Terms to be translated**
  - **Translated terms**
  - **Terminology Extraction**
A comparable corpora is taken as input
Output is a list of single- and multi-word candidate terms along with their candidate translations

Processes involved:
- Term Extraction
- Term Alignment
  - Direct Context-Vector Method
  - Translation of lexical units
Term Extraction

- Terminological units extracted are MW terms whose syntactic patterns, expressed using POS tags, correspond either to a canonical or a variation structure.
- For French main patterns are:
  - N N
  - N Prep N et N adj
- For Japanese main patterns are:
  - N N
  - N Suff
  - Adj N
  - Pref N
- Variants handled are:
  - Morphological for both French and Japanese
  - Syntactical for French
  - Compounding for Japanese
Variants handled in French

- **Morphological Variant**: Morphological modification of one of the components of base form
- **Syntactical Variant**: the insertion of another word into the components of the base form
- **Compounding Variant**: the agglutination of another word to one of the components of base form
- **Example**: sécrétion d’insuline (insulin secretion)
  - Base form: N Prep N pattern
  - Morphological variant: sécrétions d’insuline (insulin secretions)
  - Syntactic variant: sécrétion pancréatique d’insuline (pancreatic insulin secretion)
  - Syntactic variant: sécrétions de peptide et d’insuline (insulin and peptide secretion)
Variants handled in Japanese

- Example the MWT \text{インスリン.分泌} (insulin secretion) appears in following form:
  - Base form: N N pattern: \text{インスリン/N}_1 \text{分泌/N}_2
  - Compounding variant: agglutination of a word at the end of the base form: \text{インスリン/N}_1 \text{分泌/N}_2 \text{能力/N}_3 (insulin secretion ability)
Term Alignment

- It aligns source MWT’s with target SWT’s or MWT’s

Direct Context-Vector method:
- Collect all lexical units in the context of lexical unit ‘i’ in a window of size ‘n’ words around ‘i’
- For each lexical unit of source and target language
  - Obtain a context-vector $V_i$, which gathers the set of co-occurrences units $j$ associated with the number of times that $j$ and $i$ occur together
- Normalize context vector
  - Mutual information
  - Log-likelihood
Term Alignment: Direct Context-Vector Method

- Using Bilingual Dictionary, translate the lexical units of the source context-vector.
- For a word to be translated, compute the similarity between the translated context-vector and all target vectors through vector distance.
- Candidate translations of a lexical units are the target lexical units closest to the translated context-vector acc. to the vector distance.
Translation of lexical units
- Depends on the coverage of bilingual dictionary
- If bilingual dictionary provides several translations for a lexical unit, consider all of them but weight the different translations by their frequency in the target language
- For a MW, possible translations are generated by using compositional method
- If it is not possible to translate all compositions of MW, MWT is not taken into account in the translation process
Composition methods for French and Japanese

• For Japanese
  - Fatigue chronique (chronic fatigue)
    - for fatigue four translations are possible: 疲れ, 疲労, 倦怠, 飽き
    - two translations for chronique: 記事番組, 慢性
  - We generate all combinations\(^2\) of translated elements and select those which refer to an existing MWT in the target language

• For French
  - For a multi-word of length n\(^3\), produce all the combinations of MW Unit elements of length less than or equal to n
  - Syndrome de fatigue chronique (chronic fatigue disease) yields the four possible combinations:
    - [Syndrome de fatigue chronique]
    - [Syndrome de fatigue] [chronique]
    - [Syndrome] [fatigue chronique]
    - [Syndrome] [fatigue] [chronique]
  - A direct translation of subpart of the MW is done if present in bilingual dictionary
  - 90% of the candidate terms provided by term extraction are composed of only two content words, so limiting to the combination 4th

2: Grefenstette, 1999
3: Robitaille et al., 2006
Comparable Corpora

- Are sets of texts in different languages, that are not translations of each other
- Share some characteristics or features: topic, period, media, author, discourse
- One of the clearest is ICE -- the International Corpus of English Corpora of around one million words in each of many varieties of English around the world
A bilingual dictionary or translation dictionary is a specialized dictionary used to translate words or phrases from one language to another.

Bilingual dictionaries can be

- Unidirectional, meaning that they list the meanings of words of one language in another
- Bidirectional, allowing translation to and from both languages.
Conclusion

- More frequent a term and its translation, the better is the quality of alignment
  - The discourse categorization of documents allows lexical acquisition to increase precision
  - Including discourse, results in candidate translations of better quality even if the corpus size is reduced by half
  - Gives rise to data sparsity problem
- Data sparsity problem can be partially solved by using comparable corpora of high quality
References

- Bilingual terminology mining - using brain, not brawn comparable corpora [Emmanuel Morin, B´eatrice Daille, Koichi Takeuchi, and Kyo Kageura 2007]
- Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain X. Saralegi, I. San Vicente, A. Gurrutxaga