

Extracting bilingual terminologies from comparable corpora

By: Ahmet Aker, Monica Paramita, Robert Gaizauskas

CS671: Natural Language Processing
Prof. Amitabha Mukerjee

Presented By:
Ankit Modi (10104)



Introduction

- » Bilingual terminologies are important for various applications of human language technologies
- » Earlier studies may be distinguished by whether they work on parallel or comparable corpora
- » Focus on Comparable corpora is crucial as Parallel corpora is tough to find for all language pairs



Task

To extract bilingual terminologies from comparable
Corpora



Task

To extract bilingual terminologies from comparable Corpora

Comparable corpora:

Collection of source-target language document pairs that are not direct translations but topically related.



Method

- » Pair each term extracted from S with each term extracted from T

Term: Contiguous sequence of words
(No particular syntactic restriction)



Method

- » Pair each term extracted from S with each term extracted from T
- » Treat term alignment as a binary classification task



Method

- » Pair each term extracted from S with each term extracted from T
- » Treat term alignment as a binary classification task
- » Extract features for each S - T potential term pair

Decide whether to classify it as term equivalent or not (SVM binary classifier with linear kernel)



Feature Extraction

» Dictionary Based Features

1. *isFirstWordTranslated* (Binary Feature)
2. *isLastWordTranslated*
3. *percentageOfTranslatedWord*
4. *percentageOfNotTranslatedWords*



Feature Extraction

» Dictionary Based Features

5. *longestTranslatedUnitInPercentage*

6. *longestNotTranslatedUnitInPercentage*

7. *averagePercentageOfTranslatedWords*

» First 6 features are computed in both directions (S -> T and T -> S) .

In total, we have 13 Dictionary Based Features



Feature Extraction

» Cognate Based Features

1. Longest Common Subsequence Ratio:

Ex: LCSR ('dollar', 'dolari') = 5/6

2. Longest Common Substring Ratio:

Ex: LCSTR ('dollar', 'dolari') = 3/6

3 Dice Similarity:

Dice = $2 * \text{LCST} / (\text{len}(X) + \text{len}(Y))$



Feature Extraction

» Cognate Based Features

4. Needleman Wunsch Distance (NWD):

$$\text{NWD} = \text{LCST} / \min[\text{len}(X) + \text{len}(Y)]$$

5. Levenshtein Distance:

$$\text{LDn} = 1 - (\text{LD} / \max[\text{len}(X), \text{len}(Y)])$$

» We have 5 Cognate Based Features



Feature Extraction

» Cognate based features with term matching

Applicable to those pair of languages whose alphabets belong to a common character set

A mapping is performed from a source term to a target writing system or vice versa.

Same cognate features as previous are calculated in both directions

» We have 10 such features



Feature Extraction

» Combined Features

1. *isFirstWordCovered:*

Translation + Transliteration

2. *isLastWordCovered:*

3. *percentageOfCoverage:*

4. *percentageOfNonCoverage*

5. *difBetweenCoverageAndNonCoverage*

» Calculated in both directions - 10 Combined Features



Feature Extraction

» We have 38 features

Dictionary based features : 13

Cognate based features : 5

Cognate based features with term matching : 10

Combined features :10



Evaluation 1

- » Some positive and negative examples are created
- » Precision, recall and f-score are calculated
- » The precision score ranges from 100 to 67 percent



Evaluation 2

- » Manual Evaluation
- » Human assessors are asked to categorize each term pair into one of the following categories:

Equivalence, Inclusion, Overlap and Unrelated

- » Over 80 percent of the term pairs were assessed to be of the first category i.e. Equivalence.



Dataset

- » Training data taken from EUROVOC thesarus
- » English-German term-tagged comparable corpora for manual evaluation



Thank You



Manual Evaluation

- » *Equivalence*: Exact translation/ transliteration of each other
- » *Inclusion*: An exact translation/ transliteration of one term contained within the other
- » *Overlap*: Terms share at least one translated/ transliterated word
- » *Unrelated*: No word in either term is a translation/ transliteration of a word in other



Error

- » Error percentage was generally low

- » Reason for errors:

Existence of words with very similar spellings but completely different meanings



SVM Binary Classifier

- » Pair each term extracted from S with each term extracted from T
- » Treat term alignment as a binary classification task
- » Linear Kernel
- » Trade-off between training error and margin parameter, $c = 10$.



Future Work

- » Looking into the usefulness of the term pairs in various application scenarios such as machine translation etc

