Extracting bilingual terminologies from comparable corpora

Ahmet Aker, Monica Paramita, Robert Gaizauskasl
University of Shefeld
Review by : Ankit Modi (10104)

October 4, 2013

ABSTRACT

This paper proposes a method for extracting bilingual terminologies from comparable corpora. For training, they have taken their data for 20 language pairs from EUROVOC thesarus. They use the SVM binary classifier for classification. Further, they measure Precision, recall and F-score for each language pairs. Over 90 percent of the generated term pairs are exact or partial translations. For some languages, precision of 100 percent was also recorded.

1 Summary

A comparable corpora has been defined as collections of source-target language document pairs that are topically related. A term has been defined as contiguous sequences of words in both source and target languages.
Let S denote the source language document and T denote the target language document. At first each term extracted from S is aligned with each term extracted from T. For each such potential S-term and T-term pair, certain features are extracted and then SVM binary classifier is used to classify them as term equivalent or not. A linear kernel is used with the SVM binary classifier and the trade-ff between training error and margin parameter is set to be c = 10.

Features are classified into following categories:

1.1 Dictionary based features

1. **isFirstWordTranslated:** It is a binary feature which indicates if the first word (or it's initial prefix) in the S-term is a translation of the first word in T-term.
2. **isLastWordTranslated:** It is a binary feature which indicates if the last word (or it's last suffix) in the S-term is a translation of the first word in T-term.
3. **percentageOfTranslatedWords:**It indicates the percentage of words in S which have their translation in T.
4. **percentageOfNotTranslatedWords:**It indicates the percentage of words in S which have no translation inT.

5. **longestTranslatedUnitInPercentage:** Expressed as a percentage, it indicates the ratio of the number of words within the longest contiguous sequence of source words which has a translation in the target term to the length of the source term.
6. **longestNotTranslatedUnitInPercentage:** Expressed as a percentage, it indicates the percentage of the number of words within the longest sequence of source words which have no translations in the target term.
7. **averagePercentageOfTranslatedWords:** It indicates the average of *percentageOfTranslatedWords* when calculated from S to T and T to S.

The first six features are computed in both directions i.e. S to T as well as T to S. So in total, we have 13 dictionary based features.

1.2 Cognate based features

1. **Longest Common Subsequence Ratio (LCSR):** It measures the longest common non-consecutive sequence of characters between two strings.
   LCSR(X,Y) = len[LCST(X,Y)] / max[len(X), len(Y)]
2. **Longest Common Substring Ratio (LCSTR):** It measures the longest common consecutive string of characters between two strings.
   LCSTR(X,Y) = len[LCS(X,Y)] / max[len(X), len(Y)]
3. **Dice Similarity:** dice =2 * LCST / len(X) + len(Y)
4. **Needlemann Wunsch Distance (NWD):** NWD = LCST /min[len(X) + len(Y)]
5. **Levenshtein Distance:** It measures the minimum no. of operations (insertion, deletion and substitution) required to deduce one string from another. It is normalized by the following formula before being used
   LDn = 1 - (LD / max[len(X), len(Y)])

We have 5 cognate based features.

1.2.1 Cognate based features with term matching

These features are 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. After mapping, the same cognate features as mentioned above are applied, only this time in both the directions i.e. S to t as well as T to S. So we have 10 Cognate based features with term matching in total.

1.3 Combined features

1. **isFirstWordCovered:** It is a binary feature which indicates if the first word in the S-term has a translation or transliteration in the target term.
2. **isLastWordCovered:** It is a binary feature which indicates if the last word in the S-term has a translation or transliteration in the target term.

3. **percentageOfCoverage:**It indicates the percentage of source term words which have a translation or transliteration in the target term.

4. **percentageOfNonCoverage:**It indicates the percentage of source term words which have niether a translation nor a transliteration in the target term.

5. **difBetweenCoverageAndNonCoverage:** It indicates the difference between percentageOfCoverage and percentageOfNonCoverage.

These features are computed in both directions i.e. S to T as well as T to S. So in total, we have 10 combined features.
In total, we have 38 pairs.

2 Evaluation

The performance of the classifier was evaluated by two methods. Firstly, some positive and negative examples are created and then precision, recall and f-score are calculated. The precision score ranges from 100 to 67 percent.

Classifier performance results on EUROVOC data (P stands for precision, R for recall and F for F-measure). Each language is paired with English. The test set contains 600 positive and 1359400 negative examples.

	ET	HU	NL	DA	SV	DE	LV	FI	PT	SL	FR	IT	LT	SK	CS	RO	PL	ES	EL	BG
P	1	1	.98	1	1	.98	1	1	.7	1	1	1	.67	.81	1	1	1	1	1	1
R	.67	.72	.82	.69	.81	.77	.78	.65	.82	.66	.66	.7	.77	.84	.72	.78	.69	.8	.78	.79
F	.80	.83	.89	.81	.89	.86	.87	.78	.75	.79	.79	.82	.71	.91	.83	.87	.81	.88	.87	.88

Secondly, manual evaluation is done. In it, human assesors 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*.

Results of the EN-DE manual evaluation by two annotators. Numbers reported per category are percentages.

Domain	Ann.	1	2	3	4
IT	P1	81	6	6	7
	P2	83	7	7	3
Automotive	P1	66	12	16	6
	P2	60	15	16	9

Though the error percentage was low, still the main reason of those were found to be the existence of words with very similar spellings but completely different meanings

3 Future Work

Further research can be carried out to look into the usefulness of the term pairs in various application scenarios such as machine translation etc.