ATLANTIS PRESS *Advances in Social Science, Education and Humanities Research (ASSEHR), volume 122*
45th International Philological Conference (IPC 2016)

Выявление терминов-кандидатов для многоязычного терминологического словаря

Пивоварова Светлана Сергеевна[1] Захаров Виктор Павлович[2]

[1]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия
[2]Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия

Extracting Term Candidates for Multilingual (English-French-Russian) Glossary of Terms

Pivovarova Svetlana[1] Zakharov Victor[2]

[1]Saint Petersburg State University (SPSU), St. Petersburg, Russia
[2]Saint Petersburg State University (SPSU), St. Petersburg, Russia

Аннотация

Целью исследования является разработка методики извлечения терминов-кандидатов для многоязычного терминологического словаря. В основе разработанной методики лежит алгоритм extract–then–align. Для проверки алгоритма используется параллельный англо-франко-русский корпус текстов.

Ключевые слова: переводная терминология, параллельные корпусы текстов, выделение терминов

Abstract

Our research is aimed at developing a procedure for extracting term candidates for a multilingual glossary of terms. The procedure is based on the extract–then–align algorithm. The performance of the algorithm is evaluated on the parallel English-French-Russian corpus.

Keywords: translation of terminology, parallel corpora, extraction of terminology

Introduction

Progress in science and technology brings about changes in terminology of various fields of knowledge, but it often happens that new terms are not promptly registered in term bases and glossaries. It pays, therefore, to develop methods for term extraction, which are essential to lexicography and automatic creation of dictionaries, as well as to machine translation. Theoretical and practical

issues of term extraction have been widely discussed in a large number of scientific works, among which [7].

Our research is aimed at developing a procedure for extracting term candidates from a parallel corpus for a multilingual (English-French-Russian) glossary of terms. In order to achieve this goal we have first identified the main characteristics of English, French and Russian terms and their translation equivalents in the field of humanities; second, we have developed an algorithm that extracts term candidates from a multilingual parallel corpus; and finally we have applied the algorithm to the English-French-Russian parallel corpus of international legal documents.

Methods and techniques

The procedure for extracting term candidates from a multilingual parallel corpus is based on the extract-then-align algorithm which consists of two main steps: (1) extracting term candidates for each language from the multilingual corpus and (2) matching translation equivalents. The extraction of term candidates is based on a hybrid approach of term recognition which combines the use of morphosyntactic patterns and statistical measures. Thus, statistical measures are calculated for a list of term candidates previously extracted from a corpus by means of morphosyntactic filters. Matching translation equivalents draws upon a pivot language based approach which implies that terms of two languages are translation equivalents if they are translations of the same term of a third language (pivot, or intermediary, language).

Morphosyntactic patterns

In order to form a set of morphosyntactic patterns underlying English, French and Russian terms and their translation equivalents in the field of humanities we have analysed the Defence Reform Terminological Database [3] and the Dictionary of Economics and Commerce [2]. We provide analysis of one-word terms and terms of length 2 as they constitute the majority of all terms.

One-word terms of the English, French and Russian languages are nouns. Most English, French and Russian terms of length 2 have the following structure (N stands for noun, Adj — adjective, g — genitive case, a — accusative, d — dative, prep — preposition):

- English terms of length 2:
  - N + N: *army force, barter agreement*;
  - Adj + N: *wholesale trade, social work*;
  - N + of + N: *course of action, date of maturity*;
- French terms of length 2:
  - N + de + N: *compte d'avance, agence de publicité*;
  - N + Adj: *personne déplacée, balance déficitaire*;
  - N + N: *essence aviation, classe affaires*;
  - Adj + N: *grande formation, mauvaise récolte*;
- Russian terms of length 2:
  - Adj + N: *армейская группировка, расчетная группа*;
  - N + Ng: *зона конфликта, модернизация вооружения*;
  - N + prep + Na: *введение в строй, инвестиции в оборону*;
  - N + prep + Nd: *затраты по выбытию, проценты по займам*.

The analysis of the above-mentioned dictionaries enabled us to identify the following matches between pairs of terms:

- English — French:
  - Adj + N => N + Adj: *operational activities – activités opérationnelles*;
  - N + N => N + de + N: *advance account – compte d'avance*;
  - N + of + N => N + de + N: *disposition of forces – positionnement des forces*;
  - N + N => N + Adj: *shipping agent – agent maritime*;
  - N + N => N + prep + N: *border security – sécurité aux frontières*;
  - Adj + N => N + de + N: *territorial defence – défense du territoire*;
  - N + N => N + N: *host nation – pays hôte*;
- French — Russian:
  - N + Adj => Adj + N: *législation administrative – административное законодательство*;
  - N + de + N => N + Ng: *déclaration de guerre – объявление войны*;
  - N + de + N => Adj + N: *disponibilité des équipements – техническая оснащенность*;
  - N + prep + N => Adj + N: *entraînement au combat – боевая подготовка*;
  - N + N => Adj + N: *valeurs vedettes – главная ценность*;
  - Adj + N => Adj + N: *libre circulation – свободное обращение*;
  - N + prep + N => N + prep + Na: *mise en service – введение в строй*;
- English — Russian:
  - Adj + N => Adj + N: *bleak market – бесперспективный рынок*;
  - N + N => Adj + N: *exchange rate – обменный курс*;
  - N + N => N + Ng: *food supplies – поставки продовольствия*;
  - N + of + N => N + Ng: *breach of the law – нарушение закона*;
  - N + N => N + prep + Nd: *sickness benefit – пособие по болезни*;
  - N + N => N + prep + Na: *profits tax – налог на прибыль*.

Thus, morphosyntactic patterns and matches between pairs of terms determined in such a way are used as a basis for extracting term candidates and their translation equivalents from a corpus.

Statistical measures

Statistical measures were applied to extracted term candidates so as to assess their statistical significance. Statistically significant one-word terms were filtered out by means of raw frequency (frequency threshold = 5), whereas the log-likelihood coefficient was used to identify significant terms of length 2 (log-likelihood threshold = 5).

Multilingual parallel corpus

For the evaluation of our algorithm we have built a parallel English-French-Russian corpus of international legal documents (official texts of NATO and the UN) of about 412,000 words. The corpus was aligned at the sentence level by LFAligner [4], whereas for word alignment we used Anymalign — a multilingual sub-sentential aligner [1]. The TreeTagger [5], a POS-tagger which shows high accuracy (over 96%) [6], was used for annotating texts with part-of-speech and lemma information.

Experimental work

Experimental work was aimed at extracting term candidates and their equivalents from our corpus. We have used several mono- and multilingual glossaries of terms and term bases on policy, economics and law to assess relevance of the extracted candidates. Precision of the algorithm for one-word term candidates is 86% (483 out of 564 candidates are relevant). The most frequent term candidates are: *security — sécurité — безопасность; alliance — alliance — союз; development — développement — развитие; state — état — государство; force — force — сила*. Precision of the algorithm for term candidates of length 2 is 60% (224 out of 373 candidates are relevant). These are term candidates of length 2 with the biggest log-likelihood coefficients: *human right— droit de l'homme — правочеловека; indigenous people — peuple autochtone — кореннойнарод; nuclear weapon — arme nucléaire — ядерноеоружие; missile defence — défense antimissile — противоракетнаяоборона; territorial integrity — intégrité territorial — территориальнаяцелостность*. It was indicated by a number of researchers that a precision of 40% is acceptable for lexicographic purposes, therefore, our algorithm can be considered quite efficient.

Conclusion

Thus, we have developed a procedure for extracting term candidates from a multilingual parallel corpus for an English-French-Russian glossary of terms. Our algorithm detects term candidates that are not yet registered in term bases and dictionaries, variants of candidates as well as rare terms. The procedure developed within our research is efficient and universal, that is, it can be used for the extraction of term candidates and their translation equivalents of various fields of knowledge. The results obtained can be used in lexicography, for compiling mono- and multilingual dictionaries and glossaries of terms, as well as in machine translation systems.

Литература

Anymalign. URL: https://anymalign.limsi.fr (accessed 19 June 2016).

Benzhamen G., Pike M. Dictionary of Economics and Commerce. Moscow, 1993. 448 p.

Defence Reform Terminological Database.URL: http://www.nato.int/docu/other/ru/2005/050715/Database.pdf (accessed 19 June 2016).

LFAligner. URL: http://sourceforge.net/projects/aligner/ (accessed 19 June 2016).

TreeTagger — a part-of-speech tagger for many languages. URL: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (accessed 19 June 2016).

Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, 1994.

Seretan V. Syntax-Based Collocation Extraction. Berlin: Springer, 2011. 222 p.

References