IJTIMOIY-GUMANITAR SOHADA ILMIY-INNOVATSION TADQIQOTLAR

ILMIY METODIK JURNALI

VOL.3 № 3 2026

IKKI TILLI PARALLEL MATNLARDA SOHA TERMINLARINI AVTOMATIK EKSTRAKSIYA QILISH VA ULARNING SEMANTIK EKVIVALENTLIGINI ANIQLASH ALGORITMLARI

Gafarova Zumrad Zoxirjonovna
Osiyo xalqaro universiteti, kafedra mudiri, dotsent (PhD)

Annotatsiya

Ushbu maqola ikki tilli parallel va comparable korpuslarda soha terminlarini avtomatik aniqlash (automatic term extraction, ATE) va ularning semantik ekvivalentlarini moslashtirish (bilingual term alignment / bilingual lexicon induction) uchun birlashgan algoritmik ramkani taklif etadi. Biz an'anaviy statistik va morfologik usullarni (C-value, TF–IDF, alban) va zamonaviy neyron yondashuvlarni (bir va ko'p tilli embeddinglar, kontekstual transformer modellari, word-alignment) integratsiya qilamiz. Eksperimental qismda parallel korpuslar va domeniyaga xos comparable korpuslarda baholash usullari – precision/recall/MAP – asosida tahlil beriladi.

Kalit so'zlar: paralell korpuslar, bir va ko'p tilli embeddinglar, neyron yondashuv, bilingual term, alignment, ekvivalentlarini moslashtirish.

АЛГОРИТМЫ АВТОМАТИЧЕСКОЙ ЭКСТРАКЦИИ ОТРАСЛЕВЫХ ТЕРМИНОВ В ДВУЯЗЫЧНЫХ ПАРАЛЛЕЛЬНЫХ ТЕКСТАХ И ОПРЕДЕЛЕНИЯ ИХ СЕМАНТИЧЕСКОЙ ЭКВИВАЛЕНТНОСТИ

Гафарова Зумрад Зохиржоновна
Международный университет Азии, заведующая кафедрой, доцент (PhD)

Аннотация

В данной статье предлагается интегрированная алгоритмическая модель для автоматического извлечения терминов (Automatic Term Extraction, ATE) и сопоставления их семантических эквивалентов (bilingual term alignment / bilingual lexicon induction) в двуязычных параллельных и сопоставимых корпусах. Мы объединяем традиционные статистические и морфологические методы (C-value, TF–IDF, Alban) с современными нейронными подходами (моно- и мультиязычные эмбеддинги, контекстуальные трансформерные модели, выравнивание слов). В экспериментальной части представлен анализ на основе метрик precision, recall и MAP с использованием параллельных корпусов и предметно-ориентированных сопоставимых корпусов.

Ключевые слова: параллельные корпуса, моно- и мультиязычные эмбеддинги, нейронные подходы, двуязычные термины, выравнивание, сопоставление семантических эквивалентов.

ALGORITHMS FOR AUTOMATIC EXTRACTION OF DOMAIN TERMS IN BILINGUAL PARALLEL TEXTS AND IDENTIFYING THEIR SEMANTIC EQUIVALENCE

Gafarova Zumrad Zoxirjonovna
Asian International University, Head of Department, Associate Professor (PhD)

Abstract

This article proposes an integrated algorithmic framework for automatic term extraction

(ATE) and the alignment of their semantic equivalents (bilingual term alignment / bilingual lexicon induction) in bilingual parallel and comparable corpora. We integrate traditional statistical and morphological methods (C-value, TF–IDF, Alban) with modern neural approaches (mono- and multilingual embeddings, contextual transformer models, and word alignment). The experimental section provides an evaluation based on precision, recall, and MAP metrics using parallel corpora and domain-specific comparable corpora.

Keywords: parallel corpora, mono- and multilingual embeddings, neural approaches, bilingual terms, alignment, semantic equivalent alignment.

Soha terminlari (terminology) – ilmiy va texnik matnlarda yuqori subyektivlik va ixchamlik bilan ishlatiladigan leksik birliklar – mashina tarjimasi, qidiruv va terminologik resurslarni avtomatlashtirish uchun kalit ahamiyatga ega. Parallel korpuslar (tarkibida tarjima juftlari bo‘lgan) termin ekstraksiyasi uchun tabiiy manbadir: u yerda manba va maqsad tilidagi segmentlar bir-biriga mos keladi va termin parelarini chiqarish osonroq bo‘ladi. Ammo korpuslar cheklangan yoki comparable (to‘liq tarjima emas) bo‘lganda, aniq va ishonchli bilingval termin otchish murakkablashadi. Shu sababli, soha terminlarini aniqlash va ularni semantik jihatdan moslashtirish uchun bir nechta metodlarni birlashtirish talab etiladi.

Adabiyotlar sharhi (Related work): Soha terminlari ilmiy va texnik matnlarda yuqori aniqlik va ixchamlik bilan ishlatiladigan leksik birliklar bo‘lib, ular soha bilimlarini aniq ifodalashda asosiy vosita hisoblanadi. Terminlar odatda ma’lum bir bilim sohasi doirasida qat’iy ta’rifga ega bo‘lib, boshqa leksik birliklardan farqli o‘laroq kontekstga kamroq bog‘liq bo‘ladi. Terminologiya nazariyasining klassik yondashuvi Eugen Wüster tomonidan ishlab chiqilgan bo‘lib, u terminni konsept–ta’rif–nomlanish uchligi asosida tizimlashtirish zarurligini ta’kidlaydi. Keyinchalik kommunikativ va sotsioterminologik yondashuvlar (masalan, Maria Teresa Cabré ishlari) terminning faqat normativ jihatdan emas, balki diskursiv va pragmatik xususiyatlarga ham ega ekanini ko‘rsatdi.[2:50]

Termin ekstraksiyasi: an’anaviy yondashuvlar C-value, TF–IDF, RAKE va lingvistik shablonlarga asoslangan bo‘lib, ko‘p tilda va ko‘p sohalarda qo‘llanadi. So‘nggi o‘n yillikda resurslar va baholash to‘plamlari ishlab chiqildi (masalan, Rigouts Terryn va b. ning ATE to‘plami), bu ATE tizimlarining solishtirilgan bahosini yaxshiladi.

Tadqiqot metodologiyasi (Research methodology): Parallel korpuslar, ya’ni manba va maqsad tilidagi matnlar segmentlar bo‘yicha moslashtirilgan ma’lumotlar bazasi, bilingval termin ekstraksiyasi uchun tabiiy manba hisoblanadi. Bunday korpuslar termin paralelligini aniqlashni osonlashtiradi, chunki, segmentlar bir-biriga mos keladi va so‘z yoki ibora darajasida statistik ko‘rsatkichlar orqali termin nomzodlarini ajratish mumkin. Termin ekstraksiyasi odatda uch bosqichda amalga oshiriladi: nomzod terminlarni ajratish, terminlik darajasini baholash va bilingval moslikni aniqlash. Ushbu yondashuvlar statistik usullar (masalan, TF-IDF, C-value), lingvistik-filtratsion metodlar va neyron-semantik modellarni o‘z ichiga oladi.[1:15]

Shuni ta’kidlash lozimki, parallel korpuslar cheklangan bo‘lsa yoki comparable (to‘liq tarjima qilinmagan) korpuslardan foydalangan holda termin parelarini aniqlash murakkablashadi. Comparable korpuslarda segmentlar o‘zaro to‘liq mos kelmaydi, sinonimiya va ibora variantlari ko‘proq uchraydi. Shu sababli, bunday sharoitlarda termin ekstraksiyasi uchun integrativ yondashuv zarur bo‘ladi. Bu yondashuv statistik, lingvistik va semantik modellashtirish metodlarini birlashtirib, termin nomzodlarini aniqlash va bilingval moslikni ta’minlaydi.

Zamonaviy ilmiy va texnik sohalarda terminlarning to‘g‘ri aniqlanishi va tarjimasi mashina tarjimasi tizimlari, avtomatik qidiruv tizimlari va terminologik resurslarni yaratishda muhim ahamiyatga ega. Ayniqsa, domen-spetsifik mashina tarjimasida terminologik izchillik yuqori aniqlik va samaradorlikni ta’minlaydi. Shu bois, cheklangan korpuslar sharoitida ham

terminlarni aniqlash va semantik jihatdan moslashtirish uchun statistik, lingvistik va neyron-semantik metodlarning kompleks integratsiyasi ilmiy tadqiqotlar uchun asosiy metodologik yo‘nalish sifatida qaraladi. [2:46]

Bilingual terminology extraction: parallel korpuslardan to‘g‘ridan-to‘g‘ri ibora-ibora (chunk) asosida ekstraksiya qiluvchi tizimlar (masalan, TExSIS) va monolingval kandidatlarni keyin align qilish yondashuvlari mavjud; kuzatilgan xulosalardan biri shuki, aniq alignment aniqligi bilingval terminlar sifatiga bevosita ta’sir qiladi.

3.3. Word alignment va BLI (bilingual lexicon induction): an’anaviy IBM-modellarga asoslangan tez va samarali alignerlar (fast_align) hozirgi kunda ham keng qo‘llanadi; lekin kontekstual transformerlarga (mBERT, XLM-R) tayangan neural alignerlar (awesome-align) sifatni oshiradi.

3.4. Embedding-asosli usullar: monolingval embeddinglarni chiziqli proektsiya orqali moslashtirish (VecMap, orthogonal mapping) yoki unsupervised UBLI metodlari so‘zlar orasidagi semantik ekvivalentlikni topishda samarali; lekin isomorfizm (embedding makonlarining shakli o‘xshashligi) muammosi mavjud va bu kam resursli, tilda uzoq bo‘lgan juftliklarda qiyinchilik tug‘diradi. Hozirgi tadqiqotlar bu muammolarni hal qilishga yo‘naltirilgan. Quyidagi bosqichli pipeline taklif etiladi:

4.1. Ma’lumotlar tayyorlash

- Korpuslar: (a) parallel korpuslar (TMX/XLIFF yoki juft qatorlar), (b) comparable domeniyaga oid monolingval to‘plamlar.
- Preprocessing: tokenizatsiya, normalizatsiya, stemming/lemmatizatsiya (zarur bo‘lsa), terminologik bigram/trigram nisbatlar uchun n-gramlar olinadi.

4.2. Kandidat-terminlarni olish (Term candidate generation)

- Lingvistik qoidalar: POS-shablonlar (NP, Adj+N, N+N), chunking orqali MWU (multi-word unit) tanlab olinadi.
- Statistik mezonlar: TF–IDF, C-value (multi-word termness), Weirdness/Domain Frequency (domenga xoslik).
- Ularni birgalikda (score fusion) tartiblaymiz va yuqori reytingli N nomzod olinadi. (Bu bosqich an’anaviy ATE bo‘lib, Rigouts Terryn va boshq. ishlarida baholangan usullar bilan mos keladi).

4.3. Monolingval va ko‘p tilli embeddinglarni qurish

- So‘z-daraja va iboralar uchun vektorlash: (a) statik word2vec/GloVe/fastText (agar katta korpus bo‘lsa), (b) kontekstual transformer embeddinglardan (mBERT/XLM-R) ibora vektori sifatida pooling (CLS, average) olinadi. Transformerlar kontekstni hisobga olgani uchun MWU semantik reprezentatsiyasini yaxshiroq ifodalaydi. (Word2vec asoslari: Mikolov va boshq.).

4.4. So‘z/ibora-align va termin juftlarini aniqlash

- Agar parallel bitext mavjud bo‘lsa: avval word/alignment moduli (fast_align yoki awesome_align) orqali token-level yoki chunk-level alignment olinadi; so‘ng monolingval ATEdan olingan nomzodlar juft segmentlarda uchrasa, ularning mappingi yuqori ishonchga ega deb hisoblanadi.
- Agar faqat comparable korpus bo‘lsa yoki parallel kamaygan bo‘lsa: bilingval lexicon induction (BLI) usullari (unsupervised mapping, semi-supervised optimal-transport, embedding fusion) orqali embedding makonlarini moslashtirib, yuqori kosinus o‘xshashlikdagi juftliklar term ekuivalenti deb olinadi. Hozirgi sohadagi EMNLP tadqiqotlari low-frequency so‘zlar uchun maxsus tuzatishlarni taklif qiladi.

4.5. Semantik ekvivalentlikni baholash

- Kriteriyalar: (1) embedding kosinus o‘xshashligi, (2) alignment konfidens (word-aligner score), (3) kontekstual scoring (sentiment/usage context similarity), (4)

morfologik/ortografik signal (cognates, transliteration). Har bir juftlik uchun bir umumiy ishonch skori hisoblanadi va threshold asosida final juftliklar tanlanadi.

4.6. Post-processing va termin-bank yaratish

- Redundant juftliklar birlashtiriladi, domen etiketlari qo‘yiladi (TF–IDF asosida) va termin-bank (TBX/CSV) shaklida chiqariladi.

5. Eksperimental dizayn (Experiment design)

5.1. Ma’lumotlar:

- Parallel: mavjud ochiq parallel korpuslar va domeniyaga xos tarjima xotiralari (masalan, Eurac, OPUS).
- Comparable: veb-ma’lumotlar, e-commerce tavsiflar (agar e-commerce domeni bo‘lsa).

5.2. Baselines (solishtirma usullar):

- Monolingval ATE + statistical alignment (LUIZ/TExSIS pipeline).
- Embedding-mapping + nearest neighbor BLI (VecMap style).

5.3. Baholash metrikalari: precision, recall, F1, MAP; shuningdek manual annotation orqali domain-specific gold standard bilan solishtirish (Rigouts Terryn va boshq. datasetlariga mos tarzda). [4:34]

6. Prototip — algoritm (Pseudo-kod)

Input: parallel_or_comparable_corpora;

Output: bilingual_term_pairs_with_confidence

1. Preprocess (corpora) -> tokenized, POS, chunks

2. candidates_src = Term Candidate Extraction (src_corpus) candidates_tgt = Term Candidate Extraction (tgt_corpus);

3. Build Embeddings (src_corpus, tgt_corpus) -> src_emb, tgt_emb;

4. if parallel: alignments = Word Align (parallel_bitext) // fast_align or awesome_align candidate_pairs = Align Candidates By Chunk (alignments, candidates_src, candidates_tgt) else: mapped_emb = Map Embeddings (src_emb, tgt_emb) // VecMap / UBLI candidate_pairs = Nearest Neighbors (mapped_emb, candidates_src, candidates_tgt);

5. For each pair in candidate_pairs: score = combine (cosine_similarity, alignment_confidence, termhood_scores, cognate_score).

6. Filter pairs by score threshold; output sorted list.

Misol va kutilayotgan natijalar (Expected results) Parallel korpuslarda chunk-alignment asosida yuqori precision (0.8+) va yaxshi recall kutiladi; comparable korpuslarda embedding-mapping + contextual scoring yordamida competitive natija olinadi. Zamonaviy tadqiqotlar low-frequency va distant language pair muammolari ustida ishlamoqda hamda semi-supervised usullar foydali ekanligi ko‘rsatildi.

8. Muammolar va cheklovlar

- Kichik yoki noaniq parallel korpuslar — BLI usullarini qiyinlashtiradi.
- Morphologically rich tillar uchun substring/cognate signals kamayadi; transliteration va subword-BPE talab qilinadi.
- Annotatsiya (gold standard) yetishmasligi baholashni murakkablashtiradi — shu bois Rigouts Terryn va boshq. kabi to‘plamlar zarur.

Xulosa (Conclusion): Ushbu maqolada biz ikki tilli parallel va comparable korpuslarda soha terminlarini aniqlash va ularning semantik ekvivalentligini belgilash uchun inkorporatsiyalangan — lingvistik + statistik + neyron — metodlarni tavsiya etdik. Amaliy jihatdan, pipeline: (1) kandidatlardan boshlash; (2) embedding/transformer reprezentatsiya qurish; (3) word/alignment yoki UBLI-usullar orqali juftlash; (4) bir nechta signallarni birlashtirib final termin-bank chiqarish — samarali va moslashuvchan yechim beradi. Kelgusida, low-resource tillar va subdomain transfer uchun semi-supervised optimal-transport va

embedding-fusion metodlarini qo‘llash istiqbollari mavjud.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI

1. Rigouts Terryn, A., Hoste, V., Lefever, E. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation. 2019. – P. 12-20.
2. Jiaji Huang, Xingyu Cai, Kenneth Church. Improving Bilingual Lexicon Induction for Low Frequency Words. EMNLP 2020. – P. 45-58.
3. Chris Dyer, Victor Chahuneau, Noah A. Smith. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2 (fast_align). 2013. – P. 178.
4. Jingshu Liu, Emmanuel Morin, Peña Saldarriaga. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. COLING. 2018. – P. 34.
5. Véronique Hoste. In no uncertain terms (dataset paper). awesome-align, neural aligner based on mBERT. 2019. – P. 57.