FRAZEMALARNI LINGVISTIK BELGILASH USULLARI

Nurullayeva O'.

Annotatsiya. Frazeologik birliklar (idiomalar / MWE) tilning semantik-leksik qatlamida alohida o'rin tutadi va ularni aniqlash, belgilash hamda korpuslarga annotatsiyalash tilshunoslik va NLP sohalari uchun ustuvor vazifa hisoblanadi. Ushbu maqolada idiomalarni belgilashning asosiy usullari — POS-tagging, lemmalash, morfologik teglash, sintaktik parsing, semantik teglash, MWE identifikatsiyasi (IOB) hamda embedded va stand-off annotatsiya yondashuvlari — tartibli ravishda tahlil qilinadi. Har bir qatlamning vazifasi, amaliy afzalliklari va cheklovlari yoritilib, korpus loyihalarida qatlamli yondashuvni qo'llash bo'yicha tavsiyalar beriladi. Maqola korpus lingvistlari, lug'atchilar va tiltexnologlar uchun mo'ljallangan.

Kalit so'zlar: frazeologik birlik, POS-tagging, lemmalash, semantik teglash, MWE, IOB, stand-off annotatsiya.

Abstract. Phraseological units (idioms / MWEs) form a distinct semantic-lexical layer and their detection, annotation, and corpus marking are crucial for corpus linguistics and NLP. This paper surveys principal annotation techniques — POS-tagging, lemmatization, morphological tagging, syntactic parsing, semantic tagging, MWE identification (IOB), and embedded vs. stand-off annotation. The role, advantages, and limitations of each layer are discussed and practical recommendations for layered corpus annotation projects are provided. The article targets corpus linguists, lexicographers and language technology developers.

Keywords: phraseological unit, POS-tagging, lemmatization, semantic tagging, MWE, IOB, stand-off annotation.

Аннотация. Фразеологические единицы (идиомы / MWE) представляют собой особый семантико-лексический слой; их обнаружение и многослойная аннотация важны для корпусной лингвистики и NLP. В статье рассмотрены основные методы разметки — POS-теггинг, лемматизация, морфологическая и синтаксическая разметка, семантическая разметка, идентификация MWE (IOB), а также embedded и stand-off подходы. Даны практические рекомендации по многоуровневой аннотации. Целевая аудитория — корпуса лингвисты, лексикографы и разработчики языковых технологий.

Ключевые слова: фразеологическая единица, POS-теггинг, лемматизация, семантическая разметка, MWE, IOB, stand-off.

Frazeologik birliklar (idiomalar, MWE — multiword expressions) ko'pincha so'zma-so'z ma'nodan farq qiluvchi yaxlit ma'no hosil qiladi va shuning uchun ularni korpuslarda aniqlash va belgilash odatiy token-asosidagi tahlildan ko'ra murakkabdir.

Korpuslarda qatlamli annotatsiya (tokenlash → POS → lemma → morfologiya → syntactic parse → semantic tags → MWE/IOB) yondashuvi frazeologik birliklarni aniqlash, statistik tahlil va tarjima amaliyotlari uchun ishonchli poydevor yaratadi. [2]

Maqolaning maqsadi — idiomalarni lingvistik belgilashdagi asosiy usullarni tizimli bayon qilish, ularning bir-biridan farqi va birgalikda qo‘llanish imkoniyatlarini amaliy nuqtai nazardan taqdim etish.

2. Qatlamli teglashning asosiy turlari va ularning vazifalari

2.1. Tokenlash, POS-tagging va lemmalash

Tokenlash va POS-tagging — har bir tokenning nutq turkumini aniqlash jarayoni — frazeologik birlikning ichki tuzilishini aniqlash uchun zarur (masalan, tuzini oqladi: tuzini — NOUN, oqladi — VERB). Lemmalash esa barcha grammatik variantlarni bitta baza (lemma) ostida jamlaydi va avtomatik qidiruvni osonlashtiradi. [7]

Amaliy jihat: POS va lemma qatlamlari MWE-larni aniqlash algoritmlariga xomashyo beradi: ma’lum POS-shablonlar (ADJ+N, V+NP va h.k.) asosida potentsial frazealar topiladi.

2.2. Morfologik teglash. Morfologik atributlar (son, kelishik, shaxs, zamon) frazeologik birikma komponentlarining fleksiyasini tahlil qilish imkonini beradi. Masalan, tuzini — NOUN, SING, GEN; oqladi — VERB, PAST, 3SG. Bu ma’lumotlar idiomaning qaysi formalarda uchrayotganini aniqlashda zarur.

Afzallik: morfologik qatlam tarjima va shakl-moslashuvni avtomatlashtirishda yordam beradi

2.3. Sintaktik teglash (parsing) va chegara aniqlash. Sintaktik analiz frazeologik birlik ichidagi bog‘lanishlarni aniqlaydi: bog‘lovchi-munosabatlar, bosh fe’l va to‘ldiruvchi kabi. Dependency yoki constituency daraxtlari MWE chegaralarini aniqlash va birikmaning sintaktik integratsiyasini ko‘rsatishda foydali. Cheklov: ba’zi idiomalar sintaktik jihatdan fleksiyali bo‘lib, parserlar tomonidan noto‘g‘ri bo‘laklarga ajratilishi mumkin — shu sabab qo‘lda tekshirish zarur.

2.4. Semantik teglash (word sense / semantic tagging)

Semantik qatlamda token yoki fraza muayyan ma’no kategoriyasiga (masalan, majburiyat, hissiyot, vaqt, makon) bog‘lanadi. Avtomatik tizimlar (USAS va boshq.) so‘zlarga semantik kod biriktirib, soha-asosidagi tahlillarni osonlashtiradi. [3]

Amaliy imkoniyat: semantik teglash yordamida frazeologik birliklarni mavzuiy guruhlarga ajratish va korpusda tez qidiruv amalga oshirish mumkin.

2.5. MWE identifikatsiyasi va IOB teglash

MWElarni aniq belgilash uchun IOB (B-I-O) formati keng qo‘llanadi: B-MWE (boshlanish), I-MWE (ichki), O (tashqi). Masalan:

tuzini B-MWE

oqladi I-MWE

yaxshi O

IOB ketma-ketligi avtomatik modellarga frazeologik chegaralarni o‘rgatishda qulay.

2.6. Embedded vs. Stand-off annotatsiya

Embedded (inline, XML) — teglar matn ichiga joylashtiriladi, o‘qishda qulay; lekin katta korpuslarda murakkablik tug‘diradi. Stand-off annotatsiya esa teglarni matndan ajratib, alohida fayllarda (JSON, XML) saqlaydi va turli qatlamlarni parallel boshqarish imkonini beradi [6]

Tavsiyalar: katta, ko‘p qatlamli korpuslarda stand-off yondashuvi afzal; kichik loyihalar yoki taqdimot materiallari uchun embedded yetarli.

3. MWE aniqlash usullari: lug‘at + statistika + ML.

MWElarni aniqlashda bir necha yondashuv sinergiyasi eng samarali:

1. Lug‘at asosli — mavjud frazeologik lug‘atlar [1] asosida cheklangan to‘plamni aniqlash.
2. Statistik-korpus — n-gramlar, PMI (Pointwise Mutual Information), t-test va boshqa collocation o‘lchovlari yordamida potentsial MWElarni topish.
3. Mashina o‘rganish — IOB belgilari bilan o‘qitilgan CRF, BiLSTM-CRF kabi modellardan foydalanish.

Integratsiyalashgan yondashuv — avval lug‘at bilan qirqib olish, keyin statistik filtrlash va oxirida ML bilan aniqlik oshirish — amaliy loyihalarda yuqori natija beradi [5].

4. Amaliy tavsiyalar korpus loyihalari uchun.

1. Qatlamli ish oqimi: tokenlash → POS → lemma → morfologik teg → syntactic parse → semantik teglash → MWE/IOB.
2. Stand-off annotatsiya katta, bir nechta qatlamli korpuslar uchun tavsiya etiladi.
3. Semantik teglashda avtomatik tizimlarni (USAS va boshq.) qo‘llab, natijalarni lingvist-ekspert tekshiruvidan o‘tkazish zarur (Rayson et al., 2004).
4. MWE identifikatsiyasi uchun lug‘at bazasini yangilab borish va statistik-ML kombinatsiyasini qo‘llash samarali.
5. Etilgan misollar va annotatsiya qoidalari loyihadagi barcha qatnashchilar (annotatorlar) uchun aniq yozma qo‘llanma bo‘lishi lozim — bu annotatsiya ishonchliligini oshiradi.

Idiomalarni lingvistik belgilash — ko‘p qatlamli, interdisipliner jarayon bo‘lib, har bir qatlam (morfologik, sintaktik, semantik, MWE-belgilash) o‘ziga xos vazifani bajaradi. Amaliy loyihalar uchun avtomatlashtirilgan vositalar bilan ekspert tekshiruvi uyg‘unligi, stand-off yondashuvi va IOB standartlari asosida qatlamli annotatsiya eng maqbul yondashuvdir. Bu metodlar tarjima, leksik resurslar yaratish va til texnologiyalari uchun muhim baza bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Rahmatullayev Sh. O‘zbek tilining izohli frazeologik lug‘ati. Toshkent: O‘qituvchi, 1978. 389 b.
2. Rayson, P.; Archer, D.; Piao, S.; McEnery, T. The UCREL Semantic Analysis System (USAS). Lancaster: Lancaster University Research Report, 2004. 47 p.
3. Wilson, A.; Thomas, J. Semantic annotation in corpora: Applied linguistics study. Journal of Documentation. 1997. Vol. 53. Pp. 263–281.
4. Ramshaw, L. A.; Marcus, M. P. Text chunking using transformation-based learning. In: Proceedings of the Third Workshop on Very Large Corpora (ACL), 1995. Pp. 82–90.
5. Baldwin, T.; Kim, S. N. Multiword expressions. In: Indurkhya, N., Damerau, F. J. (eds.) Handbook of Natural Language Processing. 2nd ed. Boca Raton: Chapman & Hall/CRC, 2010. (Chapter).
6. Ide, N.; Suderman, K. GrAF: A graph-based format for linguistic annotations. In: Proceedings of the Linguistic Annotation Workshop / LREC, 2007. Pp. 1–8.
7. Jurafsky, D.; Martin, J. H. Speech and Language Processing. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2009.
8. Гриднева Т. В. Методика использования компонентного анализа семантики фразеологических единиц в начальной школе // Грани познания. — 2021. — № 1(72). — С. 115–123.
9. Теля В. Н. Русская фразеология: семантика, прагматика, лингвокультурология. Москва: Языки русской культуры, 1996.