81.25

USAGE OF THE FIELD INDEX IN MACHINE TRANSLATION

YENOK GRISHKYAN
Yerevan State University,
PhD in Philological Sciences
yenokgrishkyan@gmail.com
https://doi.org/10.52853/10.52853/25792903-2021.2-yguf

Abstract

The current article discusses the main problems of human and machine translations, and introduces a new lexical description in machine translation for faster and more accurate translation. The new method uses so-called field indicators or the Field Indices to facilitate the MT search engine for words by marking these words with special components creating a semantic field, and allowing the MT devices to search for the word according to its usage in the text.

The Field Index system covers the semantic description of the following main spheres: scientific field, public or social field and humanitarian field. These three spheres contain subfields that usually mingle with the parent index through a dash, with the parent index being with the first one.

The scientific field includes such aspects as geography, mathematics, chemistry, physics, economy, medicine, etc., with related subfields like diseases and biological terms (for medicine), names of drugs (as a separate filed), finance and accounting (as part of the economy), etc. Applied Sciences Index contains miscellaneous words used by other subfields of the same scientific sphere: e.g., computer, telephone, function and many others, and plays a crucial part in distinguishing polysemantic words such as mouse (hardware), root (in mathematics), etc.

The public or social field contains subfields that narrow the meaning of words to a specific one and includes aspects as art, agriculture, law, education, religion, housing utilities, time, transportation, people, etc.i+iT Ss the widest semantic field containing a lot of subfields specifying words that belong to such groups as colours, architecture, games, music, sport, etc. (for art index), clothing, beverages, food and production (for agriculture index). Notions of time, people, professions and terms for religion and items used in the household are present in this group due to its wide usage within the society.

The humanitarian sphere deals mostly with terms used in languages, literature, manuscripts and libraries. These subfields help identify polysemantic words between nations and languages, book titles and ordinary words and phrases, and literary styles (documentaries, fairy tales, dramas, etc.). In turn, these can be further defined as prose or a poem.

All formulae proposed in the project consider the presence of the Field Indices and its position at the end of the description of the word. Depending on the target language, the translated version should be identical to source following this very principle.

Keywords and phrases: machine translation, lexical meaning, metalanguage, Universal Language Theory, artificial intellect.

ԲՆԱԳԱՎԱՌԱՅԻՆ ՑՈՒՑԻՉԻ ԿԻՐԱՌՈՒՄԸ ՄԵՔԵՆԱԿԱՆ ԹԱՐԳՄԱՆՈՒԹՅԱՆ ՄԵՋ

ԵՆՈՔ ԳՐԻՇԿՅԱՆ
Երևանի պետական համալսարան
բանասիրական գիտությունների թեկնածու
yenokgrishkyan@gmail.com

Համառոտագիր

Ներկա հոդվածը համառոտակի ուսումնասիրում է մարդկային և մեքենական թարգմանության առջև ծառացած խնդիրները, ինչպես նաև ներկայացնում է մեքենական թարգմանության ընթացքում կիրառվող բառային նկարագրության նոր մեթոդ, որի շնորհիվ հնարավոր է առավել արագ ու ճշգրիտ թարգմանություններ կատարել: Առաջարկվում է բնագավառային ցուցիչը, որն անմիջապես կկցվի բառին՝ բնորոշելով նրա կիրառության ոլորտը և հստակեցնելով բառի իմաստային դաշտը: Այս ցուցիչը կհեշտացնի մեքենական համակարգի կողմից բառային ընտրությունը՝ զգալիորեն բարձրացնելով մեքենական թարգմանության որակը:

Բնագավառային ցուցիչը ներառում է հետևյալ երեք հիմնական ոլորտների իմաստաբանական նկարագրությունը՝ գիտական, հասարակական կամ սոցիալական և հումանիտար: Այս երեք ոլորտները ներառում են ենթադաշտեր, որոնք սովորաբար խառնվում են գլխավոր ցուցչին գծիկի միջոցով, որը դրվում է սկզբում:

Գիտական ոլորտը ներառում է այնպիսի ենթաոլորտներ, ինչպիսիք են աշխարհագրությունը, մաթեմատիկան, քիմիան, ֆիզիկան, տնտեսագիտությունը, բժշկությունը և այլն: Կիրառական գիտությունների ցուցիչը պարունակում է տարբեր բառեր, որոնք օգտագործվում են նույն գիտական ոլորտի այլ ենթադաշտերում՝ համակարգիչ, հեռախոս, գործառույթ և այլն, ինչն էլ վճռական դեր է խաղում բազմիմաստ բառերի տարբերակման մեջ, օրինակ՝ մկնիկ (սարքավորում), արմատ (մաթեմատիկական) և այլն:

Հասարակական կամ սոցիալական ցուցիչը պարունակում է ենթացուցիչներ, որոնք նեղացնում են բառերի իմաստը և ներառում են արվեստի, գյուղատնտեսության, իրավունքի, կրթության, կրոնի, ֆինանսների, ժամանակի, տրանսպորտի, մարդանց և այլ ոլորտներ: Սա իմաստաբանական ամենալայն ոլորտն է, որն ունի բազմաթիվ ենթադաշտեր, որոնք մակնշում են այնպիսի

խմբերին պատկանող բառեր, ինչպիսիք են ճարտարապետությունը, գույները, խաղերը, երաժշտությունը, սպորտը և այլն (արվեստի ցուցչի համար), հագուստ, խմիչքներ, սնունդ և արտադրություն (գյուղատնտեսության ցուցչի համար), իսկ ժամանակի, մարդկանց, մասնագիտությունների և կրոնի, տնային տնտեսության մեջ կիրառվող առարկաների հասկացություններն առկա են այս խմբում՝ հասարակության մեջ դրա լայն կիրառման պատճառով: Հումանիտար ոլորտը բնութագրում է լեզուներում, գրականության և մատենագրության մեջ օգտագործվող տերմինները: Այս ենթացուցիչները օգնում են բնութագրել և տարբերել լեզուները ազգերից, գրքերի վերնագրերը սովորական բառերից ու արտահայտություններից, ինչպես նաև նկարագրում են գրականության ոճերը (վավերագրական, հեքիաթային, դրամատիկ և այլն): Իրենց հերթին դրանք կարող են հետագայում սահմանվել որպես արձակ կամ բանաստեղծություն:

Ներկայացվող նախագծում բոլոր բանաձևերը տրվել են՝ հաշվի առնելով բնագավառային ցուցչի առկայությունը և նրա վերջավոր դիրքը: Կախված թարգմանվող լեզվից՝ տվյալ թարգմանված տարբերակը պետք է հետևի վերոհիշյալ բանաձևին:

Բանալի բառեր և բառակապակցություններ. մեքենական թարգմանություն, բառիմաստ, մետալեզու, համընդհանուր լեզվի տեսություն, արհեստական բանականություն:

ИСПОЛЬЗОВАНИЕ ИНДИКАТОРОВ СФЕР В МАШИННОМ ПЕРЕВОДЕ

ЕНОК ГРИШКЯН
Ереванский Государственный Университет,
кандидат филологических наук
yenokgrishkyan@gmail.com

Аннотация

В статье вкратце обсуждаются проблемы человеческого и машинного переводов, а также представляется новый метод лексического описания, используемый при машинном переводе для быстрого и точного перевода. Новый метод использует так называемые индикаторы сфер, цель которых облегчить поисковую системы слов для машинного перевода, помечая специальными компонентами, уточняя их семантическое поле. Индикаторы сфер позволят механический выбор слов, тем самым значительно повысив качество машинного перевода.

Система индикаторов сфер охватывает семантическое описание следующих основных областей: научная, общественная или социальная и гуманитарная. Эти три области содержат субсферы, которые обычно спариваются через тире с главным индексом, вставленном в начале.

Научная область включает в себя такие аспекты, как география, математика, химия, физика, экономика, медицина и т.д. с такими субсферами, как болезни и биологические термины (для медицинской сферы), названия лекарств (как отдельная сфера), финансовые и бухгалтерские термины (как часть экономической сферы) и т.д. Индикаторы прикладных наук содержат различные слова, используемые в других сферах той же научной области, например, компьютер, телефон, функция и многие другие, и играют решающую роль в различении многозначных слов, таких, как мышка (оборудование), корень (в математике) и т.д.

Общественное или социальное поле содержит субсферы, которые сужают значение слов до определенного предела, и включают такие аспекты, как искусство, сельское хозяйство, право, образование, религия, жилищно-коммунальные услуги, время, транспорт, люди и т.д. Это самое широкое семантическое поле, содержащее множество субсфер, определяющих слова, принадлежащие к таким группам, как цвета, архитектура, игры, музыка, спорт и т.д. (для сферы искусства); одежда, напитки, еда и производство (для сферы сельского хозяйства). Понятия времени, людей, профессий и терминов, обозначающих религию и предметы бытового использования, включены в эту группу из-за их широкого использования в обществе.

Гуманитарная сфера имеет дело в основном с терминами, используемыми в языках, литературе, рукописях и библиотеках. Эти субсферы помогают идентифицировать многозначные слова между народами и языками, названия книг и обычные слова и фразы, а также литературные стили (документальные, сказки, драмы и т.д.). В свою очередь, их можно определить как прозу или стихотворение.

Все формулы, предложенные в методе, учитывают наличие индекса поля и его положение в конце описания слова. В зависимости от языка перевода переведенная версия должна быть идентична исходной по этому принципу.

Ключевые слова и фразы: машинный перевод, лексическое значение слова, метаязык, универсальная теория языка, искусственный интеллект.

Introduction

For hundreds of years, the history of humanity has been through wars and conquests, and the peace after them required understanding of the conquered people: new nations that hardly understood any other language but their own. For this very reason, the position of a translator has been created and has evolved into fast, non-human automatic translating devices.

Both human (HT) and machine translations (MT) have their advantages and disadvantages. Humans demonstrate mispronunciations (due to the lack of the target language sounds or tones in the mother tongue), mistakes in the meaning of the words (possibly due to the homonyms), misinterpretation in stylistic devices (likely due to the lack of knowledge of these devices in the target language) that may result in a completely different, sometimes even opposite understanding of the real situation. However, the same humans may use feelings in translating and often quickly orientate in choosing the right words (and sometimes omitting the sentences or changing them into others to avoid bad situations). On the other hand, machine translations avoid technical problems (pronunciation, grammar, style, etc.) by using accurate translation; however, they lack this human factor of quick orientation. Machine translation has evolved and is in transition from computational devices (various translational services online and offline) to artificial intellect.

Rene Descartes idea to create a universal language with equivalent ideas in the target language sharing one symbol**[4, p 2]** was proposed in 1629, and this was later the basis of modelling a universal language able to act as a mediator, a kind of metalanguage between two foreign languages be they of the same language family or not. However, it wasn't until 1932 that the first-ever machine translation was applied by Georges Artsrouni, a French engineer of Armenian descent, who based the dictionary and the grammatical rules of the MT on Esperanto.

In 1960 the IBM manufactured its first Mark CT that aimed to solve mainly linguistic problems. The Japanese KT-1 was the first in making trilingual operations, and in 1966 the Fujitsu Company developed FACOM 230/30 English-Japanese translation. Throughout the 20^{th} century, many MT devices have been creating, surpassing each other's expectations and trying to give a better translation. Many use modelling, like those proposed by eminent Armenian academician G. B. Jahukyan: Universal Language Theory or the Japanese scientist H. Uchida’s Universal Networking Language.

All MT devices use compiled dictionaries and grammar rules when translating, while some (PROMPT, SYSTRAN, Worldlingo, ABBY Linguo) go forward by introducing choosing between aspects or fields of translation. This option becomes possible by uploading the original document in a pdf., doc(x). or txt. Extensions for

translation. Let's pause here to try and understand the benefits of choosing aspects before translating.

The most important factor in translation is the ability to express the idea of the text in the target language no matter the lexical, grammatical or stylistic devices in use, and here we run across the semantic meaning of the word. By choosing the aspect or the field of the translation before the actual process, the MT chooses from the subsystem dictionaries that contain special vocabulary. Thus, it slows down the translation process but ensures considerably higher quality.

To maximally improve the searching of correct vocabulary as well as the grammar and syntax, a field indexation is proposed that will consider the field the text belongs to and will be written in the formal description of the words according to the hierarchy making the meaning of the word much more distinct.

The proposal to use Field Index in MTs would be beneficial from the point of view of fast and accurate translation. Logically the FI is the continuation of the formal description of the word present in most, if not all MT devices. To describe the word, it is necessary to view it in its paradigmatic system, i.e. showing the word's full description, including the morphological, syntactic, and semantic structures.

Let’s take the UNL model of the word “fear” as an example and describe the word as a formula adding the FI. Due to language varieties, the paradigmatic system of nouns, as a metalanguage, generally accounts for analytical and syntactical languages, i.e. it includes declination paradigm (DP) for many I.-E. (German, Slavonic languages), Altaic, Ugro-Finnish and other language families.

fear – NDP(AN/CN/PSY)

At first, let's clarify the abbreviations. "Fear" is a noun in a declination paradigm (NDP) that is an abstract (AN) and a common noun (CN), and the most important factor is that the field to which this word belongs is psychology (SPY). Thus we have described the word "fear" in detail according to the part of speech that includes grammatical and syntactic features with FI being given in the end, and it is ready for encodification[1].

Let’s note that for decodification[2] the target language has to add its description to the word. For example, an Armenian of this decodified formula would look like this:

վախ – NDP(AN/CN/EDI_1/PSY)

There is a great deal of similarity between both formulae; however, on a closer look, the EDI_1 can be noticed within the noun declination paradigm, which stands for the formulized description of the Armenian declination system where EDI is the

[1] Encodification is the word of a certain language already formulized and described, ready to become a part of metalanguage and be translated into another language.

[2] Decodification is the translation result when the target language gives its grammatical and syntactical aspects to the word before showing in its translation.

external declination inflexion system and the subscript 1 indicates the ending into *-ի* happening to certain words in all seven cases of the declension system of Armenian.[3]

Eminent Armenian linguist L. S. Hovsepyan has written numerous articles on the formal description of Armenian in particular and UNL, in general, to be used in MT, and the present article is partially based on her proposed description of Armenian grammar to complete with the UNL English version section **[1, pp. 144-156]**.

Thus in Armenian, the equivalent word has nearly the same attributes as in English. Only inflexion is added right before the FI, i.e. it is a common abstract noun inclined by external -ի declination that belongs to the psychological field of study. The presented project demonstrates the full description system, including FI its end, so, depending on the target language, the decodification should follow this system. The formula can have the following format for monolingual and multilingual systems. The example of the multilingual formula is for Armenian, French, Spanish and Latin.

Multilingual system sample:

{տուն[NDP(MN/CN/IDI$_1$/ACN)];house[NDP(MN/CN/ACN)];maison[NDP(MN/CN/FEM/ACN)];
casa[NDP(MN/CN/FEM/ACN)];domus[NDP(MN/CN/FEM/EDI$_2$[4]/ACN)]}

Let's discuss the yet unknown abbreviations according to languages. The Armenian language, besides the common descriptions, uses MN for materialized noun and IDI$_1$ denoting that the Armenian word տուն for house belongs to the internal declination inflexion of the 1st type, which is ու> ա, which is Nom: տուն>Gen: տան (the rest of the inflexion uses the same format). Here the ACN is the FI for architecture and construction.

Filed Indices should be divided into three main groups: scientific (including physics (PYS), chemistry (CHM), biology (BIO), etc.), public/social (including art (ART), military (MLT), religion (RLG), etc.) and humanitarian (languages (LNG), literature (LIT), bibliography (BIB), etc.). Even so, we can have branches, and sub-branches of the groups, e.g. the humanitarian field includes art (ART) that uses colours (CLR); in this case, the FE is written with a hyphen (-): **red** ADJ (QLA/ART-CLR)[5].

Let's take a look at a shortlist of field indices:

Scientific field

This field mainly includes words with one sole meaning and neologisms used generally in science resulting from new areas in science. Words commonly used in medicine, geography, astrology, biology, physics, chemical elements, etc., mainly belong to this group. Here are a few examples:

AEF-Accounting, Economy, Finance

[3] Classification of the inflexion is in the UNL Armenian section manual.
[4] For Latin EDI$_2$ corresponds to the 2nd declension.
[5] QLA=qualitative adjective

ASI-Applied Sciences (IT, Math's, etc.)
CHM-Chemistry
BTN-Botany
GEO-Geography
PYS-Physics

ZOL-Zoology

APB-Amphibians
AVE-Birds
FSH-Fish
MAM-Mammals
INS-Insects
RPT-Reptiles;

Description example[6]:

Ag NDP(MN/CN/EDI$_1$/CHM)
qnpun NDP(MN/CN/EDI$_1$/ZOL-APB)

Like in the previous example, here also, we might need to use the hyphen (-) a couple of times to indicate the hierarchy: e. g. dolphin NDP (MN/CN/ZOL-MAM-FSH). The frog (qnpun) belongs to the ZOL field as an animal, but it is most known as an amphibian, that is why it is specified with APB (amphibian) description connected with the previous index through a hyphen (ZOL-APH).

Public/Social Field

This field is the widest and exceeds the other field at least twice by its indices. This is because items used by society are gradually expelled from both scientific and humanitarian fields, creating a much bigger social field. Here are some indices of the formal descriptions of the words within the public or social field. Words from journalism, architecture, items used in the house, management, military, religion, relation, people and even surnames redirect here.

DLR-Diplomacy, Law, Rights[7]
EDU-Education
FST-Festivals
JUR-Journalism (newspapers, articles, magazines redirect here)
MLT-Military
PSY-Psychology
RLG-Religion
TRS-Transportation

[6] Most examples are described for Armenian, so it is common to notice the EDI or IDI description.
[7] Quite often, similar fields might mingle and have one abbreviation.

AGR-Agriculture

FOD-Food

BVE-Drink

CLT-Clothes

ART-Art

ACN-Architecture and construction

CLR-Colors

SCL-Sculpture

DRW-Drawing

GME-Games

MSC-Music

SPR-Sport

PEO-People

NAT-Nationalities

NAM-Names

PRF-Professions

RLT_1-Relatives, close family members

RLT_2-Relatives, extended family members

RLT_3-Relatives, neighbours, friends

RLT_4-Foreigners

SNM-Surnames;

Description example:

օդաչու NDP(MN/CN/EDI_1/PEO-PRF)

Ձատիկ NDP(AN/CN/EDI_1/FST)

The use of PEO (short for *people*) in the word pilot (օդաչու) means that the word indicates a human being, and here the PRF gives the notion that it is a profession.

Humanitarian Field

This field is the narrowest of the three because it describes mostly words denoting languages, books and styles. Literary and linguistic terms, book titles, fiction and non-fictions are grouped here according to their style. Here are a few examples:

LNG-Languages

BIB-Bibliography

LIT-Literature

DCU-Documentary

CME-Comedy

DMA-Drama

ENC-Encyclopedia

FLK-Folklore

FLM-Film
POE-Poems
PRO-Prose
SPE-Speech /monologue, dialogue, rhetoric speeches/
TRG-Tragedy

STY-Stories, fairy tales.

Description sample:

Կարմիր գլխարկը NDP(MN/CN/EDI$_1$/LIT-STY(PRO))
Ձմեռվա իրիկունը NDP(MN/CN/EDI$_3$/LIT-POE)

The title of the “Red Riding Hood” (Կարմիր գլխարկը) belongs to the stories and fairy tales section (LIT-TRY). However, it still carries the PRO index showing that the famous children's story is written in prose.

In the case of homonymy between a language and a nation, extra indices are used to clarify the meaning. E. g. Armenian NDP(MN/CN/EDI$_1$/PEO-NAT) -Ñ³Û

Armenian NDP(MN/PN/EDI$_1$/PEO-NAT)-the Armenian nation
Armenian NDP(MN/PN/EDI$_1$/LIN) -the Armenian Language

As it can be noticed from the example, one of the words denoting the nation (NEO-NAT) belongs to the social field, while the other shows the language that the nation speaks; thus, it uses the humanitarian field index to describe the word.

This is just a small part of a greater and complete form of Field Indices that are compiled to aid machine translation devices or artificial intellects to create much better and more comprehensive communication between various nations, carriers of different languages.

Throughout many problems discussed within the limits of computational linguistics, the most important for Armenian nowadays is to process the texts yielding better translations. All linguistic aspects of Armenian should be formalized in detail by means of a formal description that will allow Armenian to access the network.

Language models, on the other hand, can be analytic, where the required information is extracted from the text by means of dividing it into phrases, words, auxiliary particles of language for a detailed analysis; or syntactic, where the reverse process is generated (introduction of roots, prefixes, suffixes, etc. that join together to make up the word, phrase and the sentence). This is a must for computational linguistics, because each natural language is difficult by its construction and may raise concerns when encoding into and decoding from a metalanguage.

REFERENCES

1. Լ. Հովսեփյան Հայերենի քերականության ձևայնացում և համընդհանուր ցանցային լեզվի հայերեն մոդուլի մշակում, Արևելագիտության հարցեր, Եր., 2006, հ. 6, էջ 144-156.
2. Джаукян, Г. Б., Универсальная теория языка. М., 1999.
3. Robert Batchelor, The Republic of Codes, Cryptographic Theory and Scientific Networks in the Seventeenth Century The Republic of Codes (stanford.edu)
4. Uchida, H, Zhu, M., Della Senta, T., Universal Networking language, Zurich, 2005.

Հոդվածը ներկայացվել է տպագրության 03.03.2021,
ընդունվել է տպագրության 10.04.2021