Automatic smart subword segmentation for the reverse Ukrainian physical dictionary task

Maksym Vakulenko[1,2,*,†] and Vadym Slyusar[2,†]

[1] *Darmstadt University of Applied Sciences, Schoefferstrasse 3, 64295 Darmstadt, Germany*
[2] *Institute of Problems of Artificial Intelligence, Prospekt Akademika Ghlushkova 40, 03187 Kyjiv, Ukraine*

Abstract

This article introduces a novel method for tackling the reverse dictionary task, utilizing text segmentation into subwords. We focus on physical texts written in Ukrainian, dividing words into subwords that include morphemes, individual characters, and their combinations. Unlike word-level segmentation, the subword vocabulary is limited, thereby eliminating the issue of unknown lexical units. Unlike character-level segmentation, each subword retains a certain degree of semantic information, which allows for the construction of meaningful embeddings. We explore various combinations of language models using different levels of segmentation in the context of reverse dictionary development. This approach represents a significant advancement towards automating terminological work through the utilization of machine learning methods applied to terminology science. The findings enhance the linguistic capabilities of artificial intelligence, helping it to process terminology research with a human-like comprehension. Furthermore, the consideration of the Mixture of Experts (MoE) architecture is proposed to integrate both traditional word-based and innovative subword-based approaches. This hybrid method aims to leverage the strengths of both segmentation levels, thereby enhancing the performance of multimodal large language models (LLMs) in processing and understanding intricate linguistic structures.

Keywords

reverse dictionary, subword segmentation, terminology science

1. Introduction

One significant aspect of natural language processing (NLP) tasks involves the generation or prediction of text or words. Reverse dictionaries, as outlined by Hill et al. (2016) and Yan et al. (2020), hold promise in this domain, where machine-generated lexical units are proposed based on their definitions.

Within this framework, employing subwords as fundamental linguistic units offers notable advantages over conventional methods. Compared to approaches using complete words as the smallest units, utilizing subwords circumvents issues associated with unseen words, allowing for the construction of new words using an existing subword vocabulary. Unlike character-based approaches, subword employment maintains a connection to underlying semantics (Chaudhary et al., 2018; Zhang et al., 2020; Aguilar et al., 2021). Consequently, the decomposition of words into constituents has been investigated in various NLP tasks focusing on text generation, prediction, and speech recognition (Chaudhary et al., 2018; Sennrich et al., 2016; Arčan et al., 2019; Church, 2020).

MoDaST-2024: 6th International Workshop on Modern Data Science Technologies, May, 31 - June, 1, 2024, Lviv-Shatsk, Ukraine
* Corresponding author.
† These authors contributed equally.
maxvakul@gmail.com (M. Vakulenko); swadim@ukr.net (V. Slyusar)
0000-0003-0772-7950 (M. Vakulenko); 0000-0002-2912-3149 (V. Slyusar)

It is important to highlight that the prevalent byte-pair-encoding method for word segmentation, grounded in mathematical statistics, exhibits several drawbacks (Aguilar et al., 2021). Among these, the most unpleasant is its tendency to erroneously segment compound words like "electroneutral" as "electron-eu-tral" or so instead of "electro-neutral" (Church, 2020).

At the same time, one of the major difficulties in the terminology work is conditioned by the need to process huge amounts of terminological data (L'Homme, 2013) which motivates their automated processing. In particular, an important part of terminology management is the prescriptive step where, according to ISO 704 (2000:vi), the prescribed (recommended) term should be chosen or created on the basis of its definition (see Drewer and Ziegler, 2011, 164). In this sense, the process of attributing designations to concepts in terminology science corresponds to reverse dictionary task in NLP.

Such formulation of terminological (and, more generally, linguistic) tasks in terms of machine algorithms contributes to linguistic competency of an artificial personality with artificial intelligence (see Shevchenko et al., 2023, 27-29) that manifests the person's ability for human-like thinking, effective lingual communication, and the so-called "accurate report". The last is considered, in turn, a significant sign of consciousness in mammals (Seth et al., 2004).

Little work of this kind has been done heretofore on the data coming from low-resource languages such as Ukrainian. This paper aims to address this gap by employing symptomatic statistical and analytical methods from the field of terminology science. Specifically, we will present two subword vocabularies tailored to the Ukrainian language within the domain of physics based on the "Explanatory dictionary on physics" (Vakulenko and Vakulenko, 2008). The two obtained texts will contain the simple and composed segmentation into the combined and individual subwords, respectively, that is the first step towards a reverse dictionary and other NLP tasks. We will discuss also the most efficient ways to create a reverse dictionary in the field of physics and adjacent fields by means of deep learning. From a more general perspective, this paper makes a step towards linguistic competency of an artificial personality with artificial intelligence (AI) that will be able to create new terms using human-like algorithms. This way, the typical assignments of terminology science that usually require much human work, will be translated to machines with elements of a linguistically competent AI.

2. Method and material

In this study, we undertake a supervised learning task focused on creating reverse and domain-specific dictionaries, necessitating the compilation of a linguistic unit vocabulary during the pre-processing phase. As highlighted earlier, subword segmentation emerges as the most viable method for preserving semantics, in contrast to character-level analysis, and for circumventing the challenge of unknown words, as opposed to word-level scenarios.

This segmentation of Ukrainian texts relies on the set of Ukrainian morphemes (affixes) sourced from specialized dictionaries (Sikorsjka, 1995; Karpilovsjka et al., 1998; Poljugha, 2001). A curated collection comprising 2,000 Ukrainian roots, encompassing both commonly used and domain-specific units, has been manually introduced.

Our initial approach involves the utilization of individual subwords. It is important to note that subwords exhibit significant homonymy, wherein the same combination of letters may occur in different parts of distinct words with varying meanings. We anticipate that incorporating individualized subwording into the neural network will yield averaged sense embeddings, similar to those at the word level (cf. Loureiro et al., 2021, p. 388). Additionally, as an analogue to contextual embedding models for words, we will elaborate on a vocabulary of combined subwords, wherein each sense corresponds to a combination of elementary subwords, if applicable. We hypothesize that this second approach will yield a more specific neural network

output. A comparative analysis of the results obtained from the aforementioned approaches can provide insights into the extent to which neural network predictions rely on the preliminary preparation of input data.

The definitions and explanations of terms are drawn from the "Explanatory dictionary on physics" (Vakulenko and Vakulenko, 2008) which, after the removal of in-text cross-references, comprises 6,068 distinct entries. The resulting subword vocabularies contain approximately 28,000 units each. The free Microsoft transliteration tool has been utilized to facilitate automatic text segmentation based on rules embodying both approaches.

To range the predicted terms according to their applicability, we suggest using the apt term criteria formulated in a machine-friendly manner (see Vakulenko, 2024):

1. **Exactness** (the concordance between the term meaning and its morphological structure) is understood as the cosine similarity (degree of entailment) between the definition and corresponding vocabulary entry.

2. **Essentiality** (coverage of key aspects of the concept and absence of false associations) is determined as the ratio between the largest entailment degree and the second-largest degree, as taken from the dictionary explanations.

3. **Plainness** (a clear inner form of a term) is calculated as the ratio of the number of subwords in the term coinciding with the sub-words in its definition, to the total number of subwords.

4. **Derivativity** (the ability to easily create derivatives of the word) is estimated as the absence of "nnja" and "ttja" in the word ending and the ability to add subwords to the existing word stem. The transliteration is carried out according to the National transliteration standard (DSTU, 2022; see also Vakulenko, 2023b).

5. **Good sound** (the agreement with phonotactic rules) is regarded as the absence of clusters of more than two different consonants (except "str", "zdr", "spr", "zbr", "skr", "skl", "stv", "zdv", "ntr", "ndr", "ntv", "ndv"); the absence of "ngh" following with a consonant or in the word end; absence of "shr" and "zhr"; the absence of two different neighboring vowels (except the second "u"); absence of "ry", "ghy"; the absence of "bv", "bf", "pv", "pf","mf", "mv", "lr", "ljr", "ljs", "ljsh"; the absence of final consonant clusters (except "sk", "lk", "nt", "st", "stj").

6. **Systemic feature**, or systemness (reflection in the designation belonging to a particular class of concepts) is assessed as the availability of the same form among other dictionary entries resulting in meronyms or hypernyms (hyponyms).

7. **Organic nature**, or organicity (conformance with spelling and language tendencies) is evaluated as the inverse number of maximum-length subwords.

8. **Compatibility** (the ability to be combined in terminological combinations) is estimated as the valence of the term or its closest analogs, if newly coined.

9. **Unambiguity** is estimated as an inverse total number of definitions in the dictionary corresponding to the term entry.

10. **Nominativity** (as opposed to descriptive attribute) is calculated according to the formula $K_{nom} = 1/(1+n_{conj}+n_{end})$, where n_{conj} is an inverse number of conjunctions in the collocation, and n_{end} is the number of verb endings "ty", "tysja", "tysj".

11. **Brevity** is estimated as an inverse number of symbols in the term (or an inverse number of sounds).

This selection of criteria is preferable to those described previously in rules regulating terminological work. In particular, the German standard DIN 2330 (1993, 8) determines the following basic lingual requirements for terms:

exactness (Ger. *Genauigkeit*), brevity (Ger. *Knappheit*), orientation towards accepted language usage (Ger. *Orientierung am anerkannten Sprachgebrauch*), motivation (Ger. *Motiviertheit*), derivability (Ger. *Ableitbarkeit*), absence of connotations (Ger. *Konnotationsfreiheit*), speakability (Ger. *Sprechbarkeit*), linguistic correctness / logic (Ger. *sprachliche Korrektheit / Logik*), clarity

(Ger. *Eindeutigkeit*) (see Drewer and Ziegler, 2011, 173-175). For example, exactness is understood here as a complex requirement combining one-to-one correspondence between a notion and a corresponding name with motivation clarity of a term. Such complex benchmarks should be split into simple ones that has been carried out in our apt term criteria.

3. Results

The pieces of codes generating the vocabulary of simple and combined Ukrainian subwords (Phys-Ukr) have been presented in (Vakulenko, 2024).

Here is an example of the subworded text (individual subwords):

&зор&і& &по&дв&ій&н&і& [&зір&к&и& &по&дв&ій&н&і&] &астр&. @

– &фіз&ич&н&а& &си&стем&а& &з& &дв&ох& &зір&ок&, &як&і& &з&в'яз&ан&і& &сил&ами& &тяж&і&нн&я& &і& &рух&а&ють&ся& &на&в&кол&о& &с&піль&н&ого& &центр&а& &мас&.

&зор&і& &с&палах&ов&і& [&зір&к&и& &с&палах&ов&і&] &астр&. @

– &з&мін&н&і& &зор&і&, &як&і& &різ&к&о& &та& &непер&і&од&ич&н&о& &з&мін&ю&ють& &сві&й& &блиск&. &Іноді& &ц&им& &терм&ін&ом& &по&зна&ч&а&ють& &ус&і& &еволюц&і&йн&о& &молод&і& &з&мін&н&і& &зір&к&и&, &але& &част&іш&е& – &ц&е& &син&онім& &з&мін&н&их& &тип&у& &U&V& &Кит&а&. &Перш&а& &з&. &с&. &за&реєстр&ов&а&н&а& &в& &1&9&2&4&, &си&стем&а&ти&ч&ні& &до&слідж&е&нн&я& &ц&их& &зір& &про&вод&ять&ся& &з& &кін&ц&я& &4&0&-&х& &рок&ів& &Х&Х& &стол&іт&т&я&. &3&. &с&. &ма&ють& &низь&к&у& &світ&н&ість&. &Вік& &від&ом&их& &з&. &с&. &від& &1&0&5& &до& &1&0&1&0& &рок&ів&. &С&палах&ов&а& &акт&ив&н&ість& &зір&к&и& &з& &вік&ом& &з&менш&у&єть&ся&.

&з&ріджж&е&нн&я& @

&=& &с&крапл&е&нн&я&.

&з&сув& &2&, -&у& (&де&форм&ац&і&йн&ий&) @

– &най&прост&іш&а& &де&форм&ац&і&я& &тіл&а&, &з&у&мовл&ен&а& &до&тич&н&ими& &на&пруж&е&нн&ями&; &про&явл&я&єть&ся& &у& &с&по&твор&е&н&н&і& &кут&ів& &елем&ент&ар&н&их& &пара&лел&епі&п&е&ді&в&, &з& &як&их&, &мож&на& &в&важ&а&ти&, &с&клад&а&єть&ся& &тіл&о&.

&з&сув& &3&, -&у& (&ен&ерг&ет&ич&н&ий&) @

– &з&міщ&е&нн&я& &рівн&ів& &ен&ерг&і&ї& &один& &від&нос&н&о& &одн&ого&.

&з&сув& &ізо&топ&іч&н&ий& @

– &з&міщ&е&нн&я& &один& &від&нос&н&о& &одн&ого& &рівн&ів& &ен&ерг&і&ї& &та& &спектр&аль&н&их& &лін&ій& &атом&ів& &різ&н&их& &ізо&топ&ів& &одн&ого& &хім&іч&н&ого& &елем&ент&у&; &про&явл&я&єть&ся& &так&о&ж& &в& &оберт&а&льн&их& &і& &колив&а&льн&их& &спектр&ах& &молекул&, &як&і& &міст&ять& &різ&н&і& &ізо&топ&и& &одн&ого& &елем&ент&у&.

The full text of “Explanatory dictionary on physics” subworded into simple (individual) and combined (composite) subwords, is available on GitHub: https://github.com/Mova-2020/Subworded-Explanatory-Dictionary-on-Physics-/tree/main.

4. Discussion

The same character combinations may necessitate different segmentation in various words, a phenomenon that can be observed within a terminology science framework utilizing a symptomatic statistical method (Vakulenko, 2014, 19–23; Vakulenko, 2023a, 123–132). Unlike mathematical statistics, which deals with strict quantities, symptomatic statistics focuses more

on qualitative occurrences and tendencies. Consequently, segmentation based on symptomatic statistics may differ from that favored by mathematical statistics, which tends to prioritize subword division according to the most "frequent" character combinations, disregarding alternative variants. However, accounting for different combinations of subwords leads to various patterns with differing probabilities.

For example, the letter combination "abcd" may be split into "ab&cd" with a 50% probability, "a&bcd" with a 30% probability, and "abc&d" with a 20% probability. Initially, the first variant may seem preferable, but this preference can change significantly with the addition of another letter. For instance, the split "ab&cde" may have a 10% probability, leaving 90% for "abc&de".

The subword vocabulary derived from the “Explanatory dictionary on physics” (Vakulenko and Vakulenko, 2008) contains numerous such units. For instance, the formant “vys” may appear in words like “vysylaty” (‘emit’) where the first two letters belong to the prefix and the third is the initial letter of the root, as well as in “vysity” (‘hang’), where this formant represents the root. To differentiate the formant “vysl” appearing in the words “provyslyj” (‘sagging’) and “vyslanyj” (‘emitted’), we introduce the additional subword combination &vy&sl&a&n& working for the last word. Similarly, to distinguish the homonymic formants “dal” as in “dala” (‘gave’) and in “dalekyj” (‘far’), we use the subwordings &da&l&a& and &dal&ek&, respectively. The formant “ynni” may belong to the adjective “polovynni” (‘half’) containing the suffixes “yn” and “n”, and to the noun “rjabotynni” (‘ripples’) with the differing suffixes “y” and “nn”. In this case, the most detailed segmentation is provided, which enables all possible variants: &y&n&n&i&.

Moreover, the frequencies of such divisions may vary significantly depending on the domain.

Given that many terms are internationalisms, the neural network is expected to predict terms composed of international elements. To accommodate this, subwords corresponding to international roots and affixes are introduced. For example, the stem “vizualjn” (‘visual’) is segmented into &viz&u&alj&n&.

This application of the symptomatic statistical method mirrors human-generated knowledge, which is pertinent to the reverse dictionary task. On the other hand, the predictions of the neural network align with the analytical method, imbuing the methods of terminology science with a machine learning interpretation, which represents a significant step toward intelligent execution of various terminological tasks. This supervised training enables the machine to emulate human thinking processes.

The text of the “Explanatory dictionary on physics” (Vakulenko and Vakulenko, 2008) subworded based on the described vocabularies, contains on average 4-5 subwords per word and is devoid of errors such as "*electron-eutral". Terms stemming from indigenous Ukrainian roots exhibit more similarity with their explanations compared to international terms.

The practical implementation of the proposed approach consists, first of all, in training of embeddings for Ukrainian subwords (composed and simple) using transformers and other architectures.

At the same time, taking into account the significant prior work in creating vector databases within the framework of the traditional approach using word dictionaries, it is advisable to consider the combination of the proposed approach to segmentation based on morphemes with known methods of tokenization and vector embedding of whole words. This can significantly improve the performance of NLP models, including those designed for reverse dictionary creation tasks.

The concept of effective integration of traditional and proposed methods may consist of the use of various technologies covering the key stages of textual data processing.

First of all, we are talking about hybrid tokenization with the segmentation of texts simultaneously at the level of morphemes and words. This dual approach allows the language model to track both the semantic nuances provided by morphemes and the contextual

information encapsulated in full words. In some cases, especially for processing unknown words or for lexical units with less clear morphological boundaries, character-level segmentation should also be included as an additional level of subword analysis.

The next object of modification is the stage of vector embeddings, where changes can be made in three important directions:

- embedding based on morphemes, which will allow displaying the semantic and syntactic properties of vector embeddings for the entire variety of morphemes. This can be achieved by training on a large corpus of morphologically annotated texts or by adapting existing word embeddings to morphemes using subword information;
- word embedding with morphological awareness, which consists of combining the process of morpheme embedding with the formation of word embeddings, ensuring that the resulting word vectors will reflect the contribution of individual morphemes. Appropriate unification can be done using weighted averaging or based on special neural architectures trained to compose embedding morphemes into word embeddings;
- contextual embedding using language models such as bidirectional encoder representations from transformers (BERT) or its derivatives capable of generating context-sensitive embeddings. These models can be fine-tuned on morpheme-segmented text to produce embeddings sensitive to the morphological structure of words in a given context.

Tuning the architecture of the language model covers two main aspects: (i) the inclusion of morphological information in the input layer of the language model and (ii) the corresponding adaptation of the attention mechanism. For example, the large language model (LLM) input layer should be designed to accept morpheme representations alongside traditional word tokens. This can be implemented using parallel channels of input of relevant data or on the basis of a unified representation that combines information at the level of morphemes and words, for example, as part of a concatenation operation. Changes in the attention mechanism are driven by the need to allow the model to focus on relevant morphemes or word segments when predicting or generating representations. This is especially useful for tasks that depend on understanding subtle semantic differences.

The learning strategy of LLM-modified architecture is based on joint learning on morphological and semantic tasks. Training should consist of a combination of tasks that require both morphological understanding (e.g., segmentation of morphemes, marking parts of speech) and semantic tasks (e.g., recognition of word meanings, reverse dictionary entry). This prompts the model to develop its representations that are informative at both levels.

Transfer learning and fine-tuning procedures can be used to simplify the learning process with the involvement of a pre-trained embedding and a language model as a starting point, with their further refinement on the corpus of text annotated with morphological information. This approach can significantly reduce the training time and improve the performance of the language model, relying on the existing linguistic knowledge.

Specific evaluation metrics that take into account both morphological accuracy and semantic relevance can be used to evaluate training effectiveness, ensuring that the integrated approach effectively supports the NLP target tasks.

It should be noted that integrating morpheme-based segmentation with traditional tokenization and embedding methods will initially require iterative refinement based on feedback and task-specific requirements. However, thanks to the well-thought-out integration of morpheme-based segmentation with traditional NLP methods, one can hope for the creation of Ukrainian-language models that will take linguistic nuances into account and be reliably contextual. This will lead to improved LLM performance in a wide range of language understanding tasks, including but not limited to reverse dictionary creation.

Looking at the positive aspects of combining word vectors with subword or morpheme vectors in a wider range of aspects, it is important to emphasize that this can significantly improve the ability of NLP models to understand and process language. At the same time, the beneficial effect of grouping words with similar meanings into common clusters will be preserved and strengthened, which will affect the process of finding synonyms and working with language structures in several ways. In particular, semantic accuracy will improve because integrating morpheme or subword vectors with whole word vectors can help models better understand the semantic relationships between words, especially since many words share morphemes that indicate relatedness or semantic proximity. For example, words with the same prefixes or suffixes often have similar meanings or belong to the same semantic category. This can make the process of finding semantic cognates more accurate and efficient.

In addition, the use of morpheme vectors allows us to enrich the vector space by providing additional dimensions to distinguish between words that may appear similar in meaning but have differences in usage or connotation. This will allow the LLM to better navigate the nuances of language and distinguish between words with subtle differences in meaning.

Integrating morpheme vectors with whole word vectors can make the search process of synonyms, antonyms, heteronyms, and other semantically related lexical units more flexible. Through morpheme analysis, language models can identify such units not only based on complete similarity of word forms but also based on commonality of morpheme components, which can reveal a wider range of semantic relationships.

Another positive effect is the improvement of the processing of newly created words. Models that use both whole word and morpheme vectors do better with newly created or rarely used words because they can interpret their meaning based on known morphemes. This enhances the model's ability to find semantically related units and understand language even when LLMs encounter unfamiliar terms.

Thus, the integration of morpheme vectors with word vectors not only preserves the beneficial effect of grouping similar words into common cluster groups but also greatly expands the potential of NLP models for understanding and processing linguistic data. This allows us to better perceive semantic relations, enrich the vector space, and increase accuracy and flexibility when finding synonyms of words. This approach makes it possible to create deeper and more extensive language models, capable of understanding not only the surface content of the text but also the deep structure and meaning of individual language units.

The option of combining two different types of vector data at the LLM input is not the only possible solution. Another approach is to use two different LLMs independently, one focused exclusively on processing traditional word vectors and the other on embedding only subword vectors. The idea is to further combine these different architectures into one through a special merge operation. This approach using different LLMs to process traditional word vectors and embeddings of subword vectors is a new strategy for building complex NLP systems. This approach allows one to use specialized models for different aspects of language analysis and then combine their strengths to achieve better performance on specific tasks. Let’s consider its main stages in more detail.

Step 1. Preparation of two variants of language models.

A model for traditional word vectors is trained or fine-tuned for NLP tasks using standard word vector bases. It can be, for example, a BERT, a generative pre-trained transformer (GPT), a Mistral, or any other model optimized for working with full-format words and their context.

The subword vector embedding model specializes in parsing and using subword vectors, such as morphemes or character grams. This model can be adapted for a deeper understanding of the morphological structure of language and used for tasks that require more detailed linguistic analysis. Each model is trained independently to process input data in its specialized domain to

solve the tasks of classification, information summarization, semantic analysis, etc. The output of these LLM variants is vector representations or other forms of output specific to a particular task.

Step 2. Fusion of model outputs

After obtaining the results from both types of models, these results are combined using several different methods.

The simplest way to combine is to concatenate the outputs of both models into one longer vector before further processing or classification. For a more refined combination, an attention mechanism can be applied, which determines the importance of each element of the output of both models for a specific task. It is also possible to develop and train an additional layer or neural network that specializes in merging the outputs from the two models, optimizing the merging process for specific tasks.

The effectiveness of this approach depends on the ability of the fusion procedure to qualitatively integrate information from both sources. The approach of combining the conclusions from different models makes it possible to use each model taking into account its maximum advantages, providing flexibility and the possibility of deeper data analysis. At the same time, models of different sizes and different numbers of layers can be used. However, such text processing also requires careful planning and tuning of the fusion process and can increase computational costs due to the need to manage multiple models.

Table 1

The main methods to merge language models in the Mergekit framework (Goddard, 2024)

Method	Multi-Model	Uses base model
Linear (Model Soups) (Wortsman et al., 2022)	✓	✗
SLERP (**S**pherical **L**inear Int**ERP**olation)	✗	✓
Task Arithmetic (Ilharco et al., 2023)	✓	✓
TIES (**Tr**I**m**, **E**lect **S**ign & Merge) (Yadav et al., 2023)	✓	✓
DARE (**D**rop **A**nd **RE**scale) (Yu et al., 2023)	✓	✓

A more advanced option for merging different LLMs, which has been intensively developing recently, is to combine their architectures using a special Mergekit framework (Goddard, 2024). Its feature is the possibility to obtain the resulting model of the same size and the same number of parameters as in the models that were subjected to the merging procedure. The list of the main methods of this type is presented in Table 1.

Fig. 1 is an illustration of the process of combining 4 pre-trained language models into one using the DARE TIES combined method. In this way, for example, a language model of a physical-technical orientation can be implemented, if not only the physical dictionary of subwords considered above, but also a technical dictionary formed similarly would be used to train the combined models.

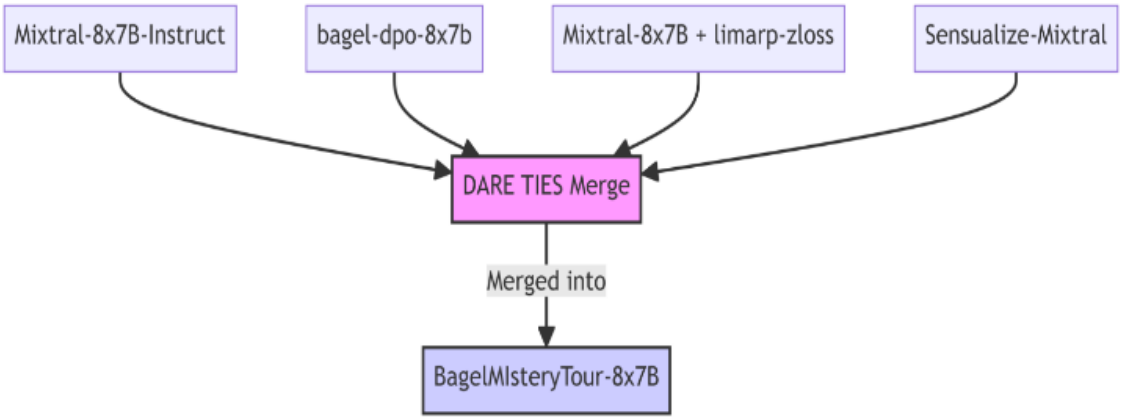

Figure 1: Merging of a few trained LLMs

An alternative option for combining models with the embedding of word vectors and subword vectors is to use the switched mixture of experts (MoE), which was developed by the team of developers of the LLM family of the Mixtral type (Mistral, 2023).

One of the first works promoting this type of architecture is the monograph by Zhi-Hua Zhou (2012). In the corresponding structure of the expert system (Fig. 2), it was assumed to control the weight vectors of the output results of several experts with the help of a special control gateway.

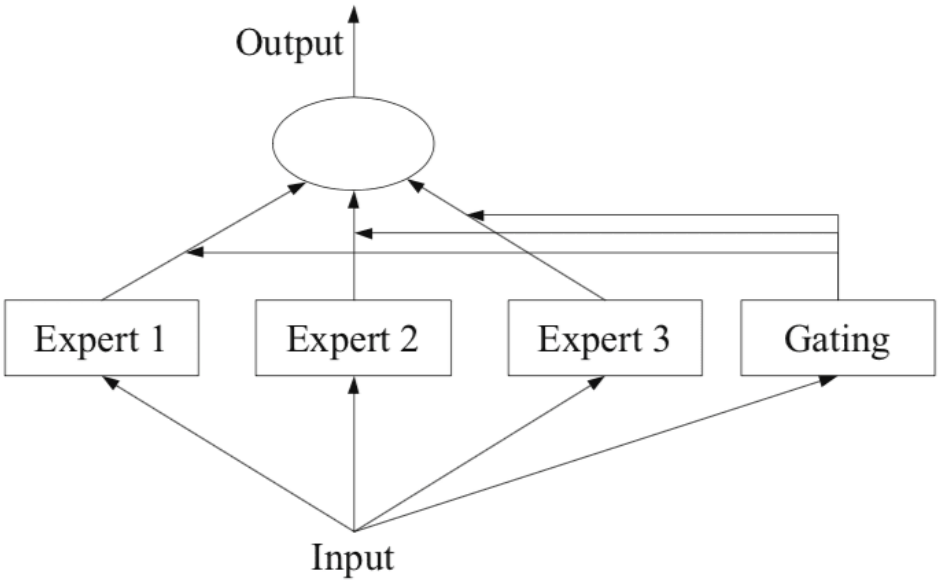

Figure 2: The classic structure of a mixture of experts (Zhou, 2012, 94)

This approach is very close to the operation of the multiple LLM merging procedure described above. In this way, the outputs of LLMs with input word embeddings and individual LLMs with subword embeddings must be combined by weight processing controlled by a special gateway.

The modern concept of MoE is an advanced approach in machine learning, which allows to create highly adaptive models by combining the conclusions from a set of “expert” subnetworks. This approach was developed in the context of LLMs such as Mixtral to improve the efficiency and adaptability of models to different tasks or data domains.

The main idea behind MoE is to distribute input data between different "expert" models based on their specialization. Each expert is optimized to handle a specific type of information or task. After processing the input data by several selected experts, the results of their work are combined using a switch that determines the weight of each expert for the final output of the model. At the same time, the rest of the experts are not involved, as shown in Fig. 3 (Chen et al., 2022), that saves computing costs and allows reducing the requirements for available hardware resources.

In the context under consideration, each of the MoE experts is proposed to be replaced by a pair of LLMs, one of which works with traditional word embedding, and the other with a vector base of subwords in the appropriate task modality.

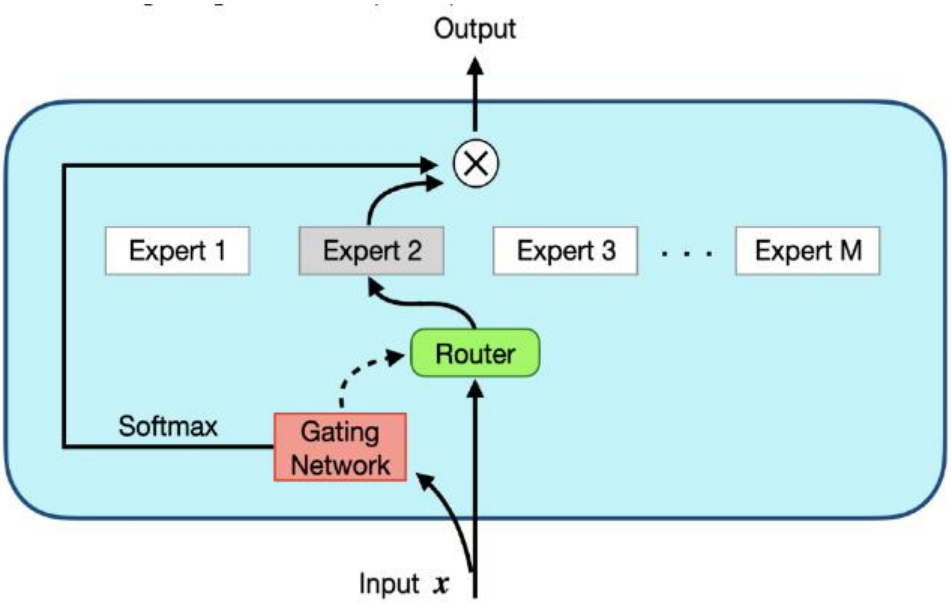

Figure 3: Switched mixture of experts (Chen et al., 2022)

The gating mechanism is also implemented on the basis of a separate language model, which decides how to distribute input data between the experts available in the structure and how to combine their conclusions. Routers can be trained to determine which expert is best to handle a given incoming request. This principle of operation allows for dynamic load distribution, adaptively changing the flow of input data between experts depending on the task or the involved context.

Thus, the MoE concept makes it possible to create models that can adapt to a variety of data types and tasks using specialized expert clusters. Adding new experts to handle additional data types or tasks is relatively straightforward, allowing for easy scaling of the model. Due to the ability to distribute the computational load among experts, MoE can be more efficient than traditional approaches, especially under resource-constrained conditions.

In the context of LLMs such as Mixtral, the MoE has been used to build models capable to efficiently handle a wide range of linguistic data and tasks, from text classification to speech generation. The MoE application option proposed by the authors makes it possible to use different expert models to process, for example, traditional word vectors and subword vectors, and then integrate their outputs to obtain a comprehensive understanding of the text. This approach opens new opportunities for the development of language models, allowing to creation more powerful, flexible, and adaptive natural language processing systems.

Using separate experts for processing words and separate experts for processing subword vectors in the context of MoE opens up opportunities to improve the flexibility and efficiency of language models and opens a possibility to involve different levels of linguistic analysis,

combining a deep understanding of the morphological structure of language with contextual analysis at the level of whole words or phrases. At the same time, expert models having various architectures, a wide range of sizes, and quantization levels can be used. This will make it possible to compensate for the increase in the volume of dictionaries of subwords compared to the traditional structures of vector bases of whole words, choosing architecture variants with a higher level of quantization of weight coefficients for the construction of expert models with morpheme embedding.

As an illustration, Fig. 4 shows the relation between the memory requirements and the number of tokens for different quantization levels (Q8, Q6 and Q5) obtained by the authors from the results of the inference procedure for LLM Dolfin 2.6 without GPU. Corresponding scores were calculated using the LM Studio framework. As expected, the memory requirements increase with the maximum number of tokens processed (horizontal axis) at all quantization levels. Significantly, such a dependence is linear, which has not been obvious. Also, higher quantization levels (Q8) require more memory than lower ones (Q6 and Q5), indicating that quantization effectively reduces memory requirements.

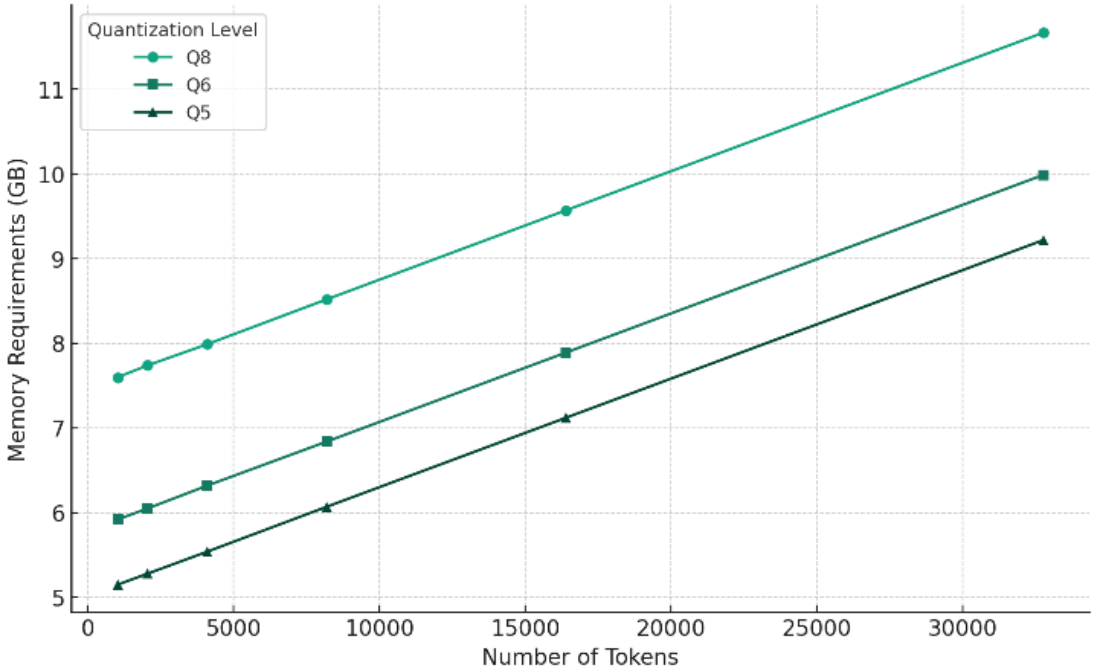

Figure 4: Memory requirements vs. number of tokens for different quantization levels

The numerical values of the data given in Fig. 4 are presented in Table 2.

Thus, LLM quantization within MoE is a key method to minimize computing resources for MoE LLM operation.

The division of tasks between experts in the MoE enables each of them to specialize in a specific aspect of language analysis. For example, whole-word experts may focus on semantic and contextual relationships, while subword experts may focus on morphological parsing and linguistic unit analysis at a finer level.

Combining input from experts specializing in different levels of linguistic analysis can lead to a deeper and more comprehensive understanding of a text. This is especially important for complex language tasks such as understanding allusions, idioms, or ambiguities.

Table 2

Memory requirements vs. number of tokens for different quantization levels

Tokens	Memory Requirements (GB)		
	8 quants (Q8)	6 quants (Q6)	5 quants (Q5)
1024	7,60	5,92	5,15
2048	7,74	6,05	5,28
4096)	7,99	6,32	5,54
8192	8,52	6,84	6,07
16384	9,57	7,89	7,12
32768	11,67	9,99	9,22

Overall, the MoE approach makes it easy to adapt the model to a variety of tasks or domains, dynamically changing the input of different experts depending on the context or specificity of the data. However, despite these advantages, training and integrating multiple specialized experts can add additional complexity to the model development and optimization process. In addition, effectively combining the findings from different experts requires careful selection and tuning of the switching mechanism to ensure an optimal distribution of weights among the experts. In doing so, it is important to ensure that no single expert dominates the decision-making process, as this may lead to insufficient consideration of input from other experts.

When scaling the considered approach to multimodal tasks, it is advisable to match image, video, or audio vectors to the embedding vectors of not only whole words but also different variants of subwords.

Similarly, in addition to the vectorization of entire images or videos, it is suggested to use a vector base of image fragments or parts of video frames. In particular, a separate augmentation of the vectorized base of video recordings by vectorizing the joints of adjacent frames in video streams can be useful, which will allow a better perception of the dynamics of interframe changes in video scenes. It is quite obvious that additional embedding of fragments or parts of video frames opens up new opportunities for deeper analysis of visual content. This is especially important for multimodal applications where visual and textual data must be matched, including embedding vectors not only for whole words but also different variants of subwords. The fact is that by analyzing individual fragments of images or parts of video frames, we can reveal details that may remain unnoticed when analyzing a complete image or video. This will provide a better understanding of the rendering scene, elements in the background, as well as smaller objects or actions that occur in the frame. Vectorization of the joints between adjacent frames allows us to more holistically and predictably perceive the dynamics of scenes, changes in the location of objects, facial expressions, or movements, providing information about the movement and interactions of all components of video content. This significantly improves the model's ability to understand video, including its verbal description.

The positive effect of multimodal interaction in the proposed way is to strengthen the correspondence between visual and textual data. In multimodal applications, it is important to establish an exact correspondence between visual elements (images, videos) and textual data (words, phrases). Vectorization of both visual and textual content at a finer level gives the model the ability to better understand the relationships between different modalities. In addition, the augmentation of the vector base due to the compatible vectorization of frame joints and subwords enriches the information space on which the model is trained, allowing it to better adapt to various tasks and contexts. This may include improving the ability to determine context, understanding intentions and emotions, and providing additional degrees of freedom for generalizations.

Although vectorizing image, audio, or video fragments increases the amount of data to process, using efficient algorithms and architectures optimized for performance can help manage this increase. At the same time, it is necessary to ensure effective coordination between different modalities, using such approaches as alignment or joint representation algorithms to integrate and synchronize vector spaces of visual and textual data. In general, the development and training of models that effectively use the extended vector base will require the use of advanced methods of deep learning and the adaptation of existing architectures to new requirements. In particular, the use of a set of small language models as part of MoE (Slyusar et al., 2024), which specializes in certain areas of combinations of subword embeddings with niche modalities, bypassing the involvement of more universal models of large sizes, deserves attention.

The use of these approaches opens up new perspectives for creating more powerful and adaptive multimodal systems that can effectively handle the complex tasks of analyzing, understanding, and generating diverse content.

Fine-tuning embeddings trained in other languages is a viable elaboration. It holds promise to benchmark the proposed method against the byte-pair-encoding technique and establish a gold standard for cosine similarity between dictionary definitions and predicted terms. Utilizing predicted terminology can augment machine translation systems, elevating translation quality. This methodology can extend to other Slavic and world languages. The created subword vocabularies can expand beyond physics to encompass various domains, including general dictionaries. Ultimately, we anticipate the development of a neural network adept at autonomously suggesting terms for emerging concepts, representing an advanced AI technology capable of performing terminological tasks. However, these pursuits necessitate dedicated investigation and computation beyond the scope of this study.

5. Conclusion

So, in this paper, we have introduced a novel method for subword segmentation essential for the pre-processing phase of reverse dictionary tasks and other natural language processing (NLP) challenges, thereby embodying the principles of terminology science within a machine learning framework. We also have established criteria for term suitability in a format compatible with machine processing, and discussed possible ways to carry out machine learning to obtain on this basis a reverse dictionary.

The resulting subworded text mitigates errors commonly encountered in widely used byte-pair-encoding algorithms, which rely solely on mathematical statistics. By employing symptomatic statistical and analytical techniques from terminology science within machine learning, we take a significant step towards executing various terminological tasks intelligently, effectively imparting human-like thinking to AI systems. Furthermore, the neural network trained to autonomously generate terms for novel concepts holds the potential to evolve into advanced AI technology capable of handling all terminological work.

References

[1] G. Aguilar, B. McCann, T. Niu, N. Rajani, N. Keskar, T. Solorio, Char2subword: Extending the Subword Embedding Space Using Robust Character Compositionality, Findings of the Association for Computational Linguistics: EMNLP (2021) 1640–1651.

[2] M. Arčan, D. Torregrosa, P. Buitelaar, Translating Terminological Expressions in Knowledge Bases with Neural Machine Translation, ArXiv 1709.02184v3 [cs.CL] (Jul 31, 2019). doi: 10.48550/arXiv.1709.02184.

[3] A. Chaudhary, C. Zhou, L. Levin, G. Neubig, D. Mortensen, J. Carbonell, Adapting Word Embeddings to New Languages with Morphological and Phonological Subword

Representations, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3285–3295. doi: 10.18653/v1/D18-1366.
[4] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, Yuanzhi Li, Towards Understanding Mixture of Experts in Deep Learning, ArXiv 2208.02813 [cs.LG] (04 August 2022). doi: 10.48550/arXiv.2208.02813.
[5] K. W. Church, Emerging Trends: Subwords, Seriously?, Natural Language Engineering 26 (2020) 375–382.
[6] Petra Drewer, Wolfgang Ziegler, Technische Dokumentation, Vogel Buchverlag, Wuerzburg, 2011.
[7] DSTU 9112:2021 (ISO 9:1995, NEQ), Kyrylychno-latynychna transliteracija i latynychno-kyrylychna retransliteracija ukrajinsjkykh tekstiv. Pravyla napysannja (Cyrillic-Latin transliteration and Latin-Cyrillic retransliteration of Ukrainian texts. Writing rules), DP UkrNDNC, Kyjiv, 2022 (in Ukrainian).
[8] Charles Goddard, Mergekit, 2024. URL: https://github.com/arcee-ai/mergekit.
[9] F. Hill, K. Cho, A. Korhonen, Y. Bengio, Learning to Understand Phrases by Embedding the Dictionary, Transactions of the Association for Computational Linguistics 4 (2016) 17–30. doi: 10.1162/tacl_a_00080.
[10] Gabriel Ilharco, Marco Tulio Riabeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, Ali Farhadi, Editing Models with Task Arithmetic, ArXiv: 2212.04089 [cs.CL] (31 Mar 2023). doi: 10.48550/arXiv.2212.04089.
[11] Je. Karpilovsjka, V. Karpilovsjkyj, N. Klymenko, T. Nedozym, Slovnyk afiksaljnykh morfem ukrajinsjkoji movy (Dictionary of affixal morphemes of the Ukrainian language), In-t movoznavstva im. O. O. Potebni, Kyjiv, 1998 (in Ukrainian).
[12] M. L'Homme, Large Terminological Databases, in: R. Gouws, U. Heid, W. Schweickard, H. Wiegand (Eds.), Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography, De Gruyter Mouton, Berlin, Boston, 2013, pp. 1480–1486.
[13] D. Loureiro, K. Rezaee, M. T. Pilehvar, J. Camacho-Collados, Analysis and Evaluation of Language Models for Word Sense Disambiguation, Computational Linguistics 47 (2) (2021) 387–443.
[14] Mistral AI team, Mixtral of experts: A high quality Sparse Mixture-of-Experts, 2023. URL: https://mistral.ai/news/mixtral-of-experts/.
[15] L. Poljugha, Slovnyk ukrajinsjkykh morfem (Dictionary of Ukrainian morphemes), Svit, Ljviv, 2001 (in Ukrainian).
[16] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, volume 1: Long Papers, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. doi: 10.18653/v1/P16-1162.
[17] Anil K. Seth, Bernard J. Baars, David B. Edelman, Criteria for consciousness in humans and other mammals, Consciousness and cognition 14 (2004) 119–139.
[18] A. Shevchenko, Y. Kondratenko, V. Slyusar, Y. Zhukov, G. Kondratenko, M. Vakulenko, Main directions for implementation of the artificial intelligence strategy in Ukraine, in: V. Vychuzhanin (Ed.), Information processing in control and decision-making systems, Problems and solutions, Odesa, Ukraine, 2023, pp. 7–33.
[19] Z. Sikorsjka, Ukrajinsjko-rosijsjkyj slovotvorchyj slovnyk: 2-ghe vyd. Slovnyk (Ukrainian-Russian word-making dictionary. 2nd edition. Dictionary), Osvita, Kyjiv, 1995 (in Ukrainian).
[20] V. I. Slyusar, Ju. P. Kondratenko, A. I. Shevchenko, T. V. Jeroshenko, Some Aspects of Artificial Intelligence Development Strategy for Mobile Technologies. Journal of Mobile Multimedia (2024). 2024, Vol. 20_3, 525–554. - doi: 10.13052/jmm1550-4646.2031.
[21] M. O. Vakulenko, O. V. Vakulenko, Tlumachnyj slovnyk iz fizyky: [6644 statti] (Explanatory dictionary on physics: [6644 articles]), VPC "Kyjivsjkyj universytet", Kyjiv, 2008 (in Ukrainian).

[22] M. Vakulenko, Term and terminology: basic approaches, definitions, and investigation methods (Eastern-European perspective), Terminology Science & Research 24 (2014) 13–28.
[23] M. O. Vakulenko, Suchasna ukrajinsjka terminologhija: metodologhija, kodyfikacija, leksykoghrafichna praktyka (Modern Ukrainian Terminology: Methodology, Codification, and Lexicographic Practice) (Specialty 10.02.01 – Ukrainian Language), Dr. Sc. thesis, Kyjiv National University after Taras Shevchenko, Kyjiv, 2023a (in Ukrainian).
[24] M. O. Vakulenko, Normalization of Ukrainian letters, numerals, and measures for natural language processing, Digital Scholarship in the Humanities 38 (3) (2023b) 1307–1321.
[25] Maksym Vakulenko, Terminology Science in Machine Learning: Smart Subword Segmentation of Ukrainian Physical Texts, in: Thomas S. Clary (Ed.), Horizons in Computer Science Research, volume 24, Nova Science Publishers, Inc., New York, 2024, pp. 147–161.
[26] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, Ludwig Schmidt, Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, ArXiv: 2203.05482 [cs.LG] (01 Jul 2022).
[27] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, Mohit Bansal, TIES-MERGING: Resolving Interference When Merging Models, ArXiv: 2306.01708 [cs.LG] (27 Oct 2023).
[28] H. Yan, X. Li, X. Qiu, B. Deng, BERT for Monolingual and Cross-Lingual Reverse Dictionary, in: T. Cohn, Y. He, and Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020 (November 2020) 4329–4338.
[29] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch, ArXiv: 2311.03099 [cs.CL] (06 Nov 2023).
[30] A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, Dive into Deep Learning, 2020. URL: https://d2l.ai/.
[31] Zhi-Hua Zhou, Ensemble Methods: Foundations and Algorithms (Chapman & Hall/CRC Machine Learning & Pattern Recognition), CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2012.